FROM CONTENT-BASED TO SEMANTIC IMAGE RETRIEVAL

Low level feature Extraction, Classification using image processing and neural networks, content based image retrieval, Hybrid low level and high level based image retrieval in the compressed DCT domain

AAMER SALEH SAHEL MOHAMED, MSc

Submitted for the degree of
Doctor of Philosophy

Department of Electronic Imaging and Media Communication

University of Bradford

2010

Abstract

Digital image archiving urgently requires advanced techniques for more efficient storage and retrieval methods because of the increasing amount of digital. Although JPEG supply systems to compress image data efficiently, the problems of how to organize the image database structure for efficient indexing and retrieval, how to index and retrieve image data from DCT compressed domain and how to interpret image data semantically are major obstacles for further development of digital image database system. In content-based image, image analysis is the primary step to extract useful information from image databases. The difficulty in content-based image retrieval is how to summarize the low-level features into high-level or semantic descriptors to facilitate the retrieval procedure. Such a shift toward a semantic visual data learning or detection of semantic objects generates an urgent need to link the low level features with semantic understanding of the observed visual information. To solve such a "semantic gap" problem, an efficient way is to develop a number of classifiers to identify the presence of semantic image components that can be connected to semantic descriptors. Among various semantic objects, the human face is a very important example, which is usually also the most significant element in many images and photos. The presence of faces can usually be correlated to specific scenes with semantic inference according to a given ontology. Therefore, face detection can be an efficient tool to annotate images for semantic descriptors. In this thesis, a paradigm to process, analyze and interpret digital images is proposed. In order to speed up access to desired images, after accessing image data, image features are presented for analysis. This analysis gives not only a structure for content-based image retrieval but also the basic units

for high-level semantic image interpretation. Finally, images are interpreted and classified into some semantic categories by semantic object detection categorization algorithm.

Acknowledgements

First of all, my strong thanks to ALLAH the most merciful; without his help and blessing, this thesis would not have progressed nor have seen the light.

I must give my deepest regards and thanks to my parents who support me with their love and encouragement. Their supplication makes me this chance to get a new life which I enjoy and makes my dream come into reality.

I would like to express my gratitude and special thanks to Professor Jianmin Jiang as primary supervisor. His expertise, thoughtful ideas, constant support, guidance, constructive suggestion and supervision helped me finish my PhD research work.

Many thanks are given to my second supervisor, Dr. Stan Ipson for his help in my study and work. He not only taught me a lot but also gave me the chance to develop my research writing skills.

Many thanks are given to my assistance supervisor, Dr. Fouad Khelifi for his help and valuable comment in my study and work. He not only taught me a lot but also gave me the chance to gain experiences in research study.

Many thanks are given to assistance researcher Dr Omer and Dr. W.Ying for their help and valuable comments in my study and work.

I would like to thank lots of my past and current colleagues, Hui, Irianto, Mely, Husam, Ibrahim,Wessam, Sukina and Jawad with whom I had happy days during my difficult times. I ought to say I would have no chance to finish my work without them.

Finally, I would like to express my thanks to my wife and four lovely children, who encourage me when I feel down and support me when I have no confidence.

Author's Contribution

Journal :-

1. Aamer Mohamed, F. Khelifi, J. Jiang, S. Ipson, "Improving Content Based Image Retrieval using Semantic Object Detection", Submitted to Journal of Image and Computing vision.

Conferences:-

1. Aamer Mohamed, Y. Wang, J. Jiang, S. Ipson "Face Detection Based On Skin Colour In Image By Neural Networks", International Conference on Intelligent and Advanced Systems, 2007, pp.779-783, Kuala Lumpur, Malaysia.

2. S. Y. Irianto, Aamer Mohamed, J. Jiang, "Integrated Keywords and Image Content features For Image Indexing and Retrieval image within Compressed Domain", Proceeding of 7th Informatics Workshop for Research.2007,211-214.Bradford , UK.

3. Aamer Mohamed, Y. Wang, J. Jiang, S. Ipson, "Face Detection Based Neural Networks Using Robust Skin Colour Segmentation", 5th International Multi Conference on Systems, Signals and Devices, 2008, Jordan.

4. Aamer Mohamed, Y. Wang, J. Jiang, S. Ipson, "An Efficient Face Image Retrieval Through DCT features", Proceeding of the 10th IASTED International Conference on Signal and Image Processing , 2008, 467-469, Hawaii , USA.

5. Aamer Mohamed, F. Khelifi, Y. Wang, J. Jiang, S. Ipson, "An Efficient Image Retrieval Through DCT Histogram Quantization", International Conference on Cyberworlds, 2009, 237-240.

6. Aamer Mohamed, F. Khelifi, J. Jiang, S. Ipson, "An Efficient Feature Extraction Techniques Based On The Histogram Of Block DCT Coefficients", Proceeding of the 7th

IASTED International Conference on Signal Processing, Pattern Recognition and Application , Feb 2010, 146-149, Innsbruck, Austria.

7. Aamer Mohamed, F. Khelifi, J. Jiang, S. Ipson, "Efficient Content Based Image Retrieval Based Retrieval Based On Semantic Object Detection", Accepted by ISSPA, 2010, Kuala Lumpur, Malaysia.

1 Table of Contents

List of figure

List of Table

List of Abbreviations

2D	Two Dimensional
ANN	Artificial Neural Network
BDIP	Block Difference of Inverse Probabilities
BVLC	Block Variation of Local Correlation coefficients
CA	Coefficient Approximation
CBIR	Content Based Image Retrieval
CCNN	Cascade Correlation Neural Networks
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
EHD	Edge Histogram Descriptor
FP	False Positive
FN	False Negative
GLCM	Grey Level Co-occurrence Matrix
HH	High High subband
HL	High Low subband
HSV	Hue Saturation Value
IR	Information Retrieval
JPEG	Joint Photographic Experts Group
KLT	Karhunen Loeve Transform
KNN	*K* - Nearest Neighbours Algorithm
LDA	Linear Discriminate Analysis
LH	Low High subband
LL	Low Low subband
ML	Maximum Likelihood
MLP	Multi Layer Persception
MPEG	Moving Picture Experts Group
MSE	Mean Square Error
NRGB	Normalized Red Green Blue
PCA	Principle Component Analysis
PDM	Point Distributions Model
QBE	Query By Example
QBIC	Query By Image Content
QBT	Query By Text
RGB	Red Green Blue
RF	Relevance Feedback
ROC	Receiver Operating Curve
SOD	Semantic Object Detection
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VQ	Vector Quantization
YCbCr	Luma(Y) Chroma blue(Cb) Chroma red(Cr)

CHAPTER ONE

1. Introduction

Recently, there has been a surge in the use of digital images with the growing availability of computers and the Internet, particularly since digital image media creation is increasing rapidly. The user demand and the availability of inexpensive storage devices as well as high quality printers allows public users to easily collect and print digital images from the Internet. Furthermore, the rapid growth of network technologies has promoted the use of digital images as one of the most important communication media for daily life.

Content-based image retrieval is the retrieval of relevant images from an image database based on automatically derived features, such as colour, texture, and shape which represent the information content of the image. The need for efficient content-based image retrieval has increased enormously in many application areas such as biomedicine, crime prevention, military, commerce, culture, education, and entertainment. In this thesis I highlight some of the work that is most related to my research. In the commercial domain, QBIC, Query by image content [1] was one of the earliest systems and was developed by IBM (www.qbic.almaden.ibm.com). Another is AMORE, a system developed by C&C NEC. Research Laboratory [2]. Besides these, there are also image search engines running on the Web such as Yahoo (www.yahoo.com), Google (www.google.com) and AltaVista (www.altavista.com). Furthermore there is new generation of online users sharing video or images over well known YouTube.

In the academic field, MIT Photobook [3] was one of the earliest systems. Columbia VisualSEEK and WebSEEK [4], and Stanford SIMPLIcity [5] are some of the more of the recent systems.

Many different indexing and retrieval methods have been used in these image retrieval systems. Some systems use keywords and full-text descriptions to index images. Others use features such as colour histogram, colour layout, local texture, wavelet coefficients, and shape to represent or describe images.

In contrast with the availability and fast development of hardware, the availability of software for managing huge image databases is not yet as capable of as users would like, in particularly for content base image retrieval (CBIR). Many types of software have been developed recently, but the results are still not satisfactory. Broadly, the problem is how to allow users access to images containing specific image contents on an image database, so the retrieved images are as expected according to the user query.

For these reasons, motivated by emergent requirements, my research aimed to investigate and propose techniques to overcome these problems. The term, "image retrieval", specifically identifies the research domain, which originates from the well developed academic discipline, Information Retrieval (IR). More particularly the term, "content-based image retrieval" (CBIR) in the DCT domain may be applied to this thesis. The development of image retrieval may involve broadly related but different areas, such as image segmentation, image feature extraction and analysis, classification, query, and retrieval.

1.1 **Background**

Recently digital images collections in many areas such as commerce, government, academia, and hospitals, are the products of digitizing existing collections of analogue photographs, diagrams, paintings, prints, etc. Image compression is a powerful tool enabling a new wave of technology that plays an important role in the use of digital information. Among various types of data commonly transferred over networks, image and video data take up over 95% of the volume on the Internet [6]. Among the several compression standards available, the JPEG [7] and MPEG [8] standards are in widespread use today. JPEG uses the Discrete Cosine Transform (DCT) applied to 8-by-8 blocks of image data. In addition, DCT preserves a set of useful properties such as energy compacting, and image data decorrelation. Thus, direct feature extraction from the DCT domain could provide better solutions for characterizing the image content, apart from its advantage in eliminating any necessity for transforming the compressed image and detecting its features in pixel domain [9]. The newer standard JPEG2000 is based on the Wavelet Transform (WT). This offers multi-resolution image analysis, which appears to be well matched to the low level characteristic of human vision. Wavelets provide a basis which is more suitable for representing images. This is because they can represent information at a variety of scales, including local contrast changes, as well as larger scale structures and thus provide a better fit for image data than non multi-resolution approaches.

The use of collection of digital images has become widespread, in both electronic and paper publications and the need for tools to manage this rapidly increasing the quantity of visual data is greater than ever. In particular, a typical image database usually has a large volume of data. Meanwhile, how to efficiently retrieve the relevant images according to users'

requests has become an urgent issue. For this purpose, Content based image retrieval (CBIR) [10] has been proposed and become one of the hot research fields in relation to image databases.

The idea is to search on an image's visual content directly. Retrieval is performed by image example where a query image is given as input by the user and an appropriate metric is used to find the best matches in the corresponding feature space.

Content-based image retrieval (CBIR) addresses the problem of finding images relevant to users' information query needs from image databases, based principally on low level visual features for which automatic extraction methods are used. As millions of images are being created daily on the WEB, image retrieval and content management becomes increasingly important across several subject areas including computer science, information systems and image processing. Since image data size is substantially larger than that of traditional text-based information, data compression is utilized to reduce the file sizes to more manageable ones without damaging the quality of images for information delivery purposes. As a result, a series of international image compression standards have been developed by worldwide organizations such as JPEG and MPEG. While image compression technology successfully resolves the problem of image file size and optimizes the use of digital storage in computers and networks, the question of how to access the content of those compressed images efficiently has become an emerging issue for further research, in order to do the feature detection or extraction for those compressed images, the conventional approaches need to decode the images to the pixel domain first, before carrying on with other existing image processing and analysis techniques [11]. This is not only time consuming, but also computationally expensive. Yet it becomes more and more important to improve the

efficiency of indexing and retrieving compressed images. Therefore, a new wave of research effort is directed at feature extraction in the DCT domain [12-14]. Visual information systems have now been developed to such a stage, that text-based content search is no longer sufficient and another new wave of research is being carried out to search target images based on their content. That is, given an example image, one wishes to search for images with similar content, which could mean similar in colour, similar in texture or similar in containing certain formulated shapes. In this context, content-based search is carried out to construct some feature keys according to the query, and then all images inside the databases are examined by a measurement of metric distance between the query feature and the features of the stored images. The first few images with the closest or with the best match between the feature vectors are retrieved as the possible candidates. This is the major research approach adopted in the community of computer vision, image processing and pattern recognition communities. Other alternative features also exist depending on how the image content is interpreted by end-users. In fact, this problem remains an unsolved issue since it is extremely hard to predict the user's intention or content interpretation. Therefore, semantic retrieval from content-based image search is an extremely difficult problem, which is now still under intensive investigation across several disciplines including software engineering, psychology and information systems.

In the area of image processing, low-level feature based approaches such as colour, texture, and shape do provide effective content extraction for physical similarity.

Human beings are likely to use high-level features (concepts), such as keywords, text descriptors, to interpret images and measure their similarity, while the image features extracted automatically using computer vision techniques mostly rely on low-level features

such as colour, texture, shape, spatial layout, etc. In general, there is no direct link between the high-level concepts and the low-level features [15]. Though many algorithms have been designed to describe colour, shape, and texture features, these algorithms cannot adequately model image semantics and have many limitations when dealing with broad content image databases [16]. Extensive experiments on CBIR systems show that low-level contents often fail to describe the high level semantic concepts in users' minds [17]. Therefore, the performance of CBIR is still far from user's expectations.

1.2 Problem definition

1.2.1 Problem 1

The difficulty of face appearances in images; faces usually vary from image to image or from person to person, depending on the lighting condition, the facial expression, the pose, and occluded objects. All these factors ensure the problem of face detection might be hard to solve.

To develop a trainable system for frontal face detection based on analyzing the skin colour of different human faces, neural networks are used as classifier to classify the presence of face and non-face in the image database.

1.2.2 Problem 2

To save the large amount of time needed to decode compressed images, compressed domain processing which refers to the DCT domain is preferable. Most of the proposed techniques are implemented directly in the DCT domain, including feature extraction, object detection, and semantic based image retrieval, which have proved to be efficient in such a context.

1.2.3 **Problem 3**

To narrow the semantic gap between low-level features and high-level semantic features, the extraction of high level features from low level features from images is a good attempt in such a context. In fact, visual saliency usually contains different objects under various patterns condition, and they can be detected through a combination of several techniques. Accordingly, these visual content objects can then be used for automatic semantic and content-based retrieval applications.

1.3 **Thesis organization**

1.3.1 **Aims and Objectives**

The primary aim of the research is to design a system, from low level feature toward high level features, or it can be stated as from content based to semantic based image database retrieval and to improve some existing techniques for content-based database management. First of all, two algorithms for efficient processing and detecting frontal face in image using neural networks will be proposed and designed; all features accessed based on the DCT domain. Then the image content will be analyzed to extract some useful features representing the image content in order to applied new proposed algorithms for image classification and retrieval. Finally, the images will be interpreted from a high-level semantic point of view. At the same time, the users should be able to index and retrieve images using both high-level and low-level features of images efficiently and effectively.

The characterization of semantic image content is meaningful to users. Using the content-based digital image system, users can query the specific semantic objects such as face images, flowers, and lion images categories. It leads to a framework that focuses on popular

indexing features such as the semantic objects, which allows people to fully exploit large amounts of image resources.

1.4 Overview of the thesis

The rest of this thesis is made up of five chapters which are organized as follows:

1.4.1 Chapter 2. Literature Review

A comprehensive literature survey is presented on face detection, content based image retrieval in the DCT domain, image compression based DCT domain and semantic based image retrieval.

1.4.2 Chapter 3. Methodology of Content based Image retrieval and classification

In this chapter, the contents of images in a database are analyzed with the proposed algorithms. By doing this, all the images can be classified and retrieved. Furthermore, image features are extracted directly from DCT domain to further process the image semantics.

1.4.3 Chapter 4. Semantic based image retrieval

In this chapter, face detection based on skin colour using neural networks from DCT domain are introduced first. Following this, an algorithm is proposed to improve the detection of faces based on robust skin colour segmentation and all algorithms are designed to access the low-resolution pixel information from DCT domain directly and efficiently.

1.4.4 Chapter 5. Hybrid semantic based image retrieval

High-level semantic image interpretation is the final destination for all the image analysis techniques. In this thesis, specific interpretation techniques are introduced. The objective of the algorithm is based on semantic object detection by using hybrid semantic based image retrieval.

1.4.5 **Chapter 6. Conclusions and future work**

The research achievements are summarized in this chapter and some plans for possible future improvements to this work are suggested.

CHAPTER TWO

2 Literature review

2.1 Survey of Face detection

In any face processing system the first step is detecting the locations in images where faces are present. However, face detection from a single image is a challenging task because of variability in scale, location, orientation (up-right, rotated), and pose (frontal, profile). Facial expression, occlusion and lighting conditions also change the overall appearance of faces.

There are many closely related problems of detection. Face localization aims to determine the image position of a single face based on the assumption that only a single face exist in an input image; Face tracking aims to estimate the location or the orientation of one or more faces in an image sequence. Face authentication aims to verify the claim of the identity of an individual in an input image, while face feature detection (face alignment) aims to locate the face features such as eyes, nose, nostrils, eyebrow, mouth, lips, ears, etc. Facial expression recognition aims to identify or detect the facial expression (sad, happy, smile, angry, and etc) of a face. Face Recognition compares the input image against others within a database and returns the closest match if any.

This survey of face detection consists of two different parts; face detection in the pixel domain and face detection in the compressed based DCT domain.

2.1.1 Face detection in the pixel domain

Face detection algorithms in the pixel domain can be classified into four groups which are reviewed in the following subsections.

2.1.1.1 **Feature based approaches**

A typical face pattern is considered to be a set of facial features such as eyes, mouth and nose; with positions and sizes within an oval shaped area. Low level feature analysis involves dealing with the segmentation of the visual features using pixel properties such as grey-scale and colour. Due to the low level nature, features generated from this analysis are ambiguous. These visual features in feature analysis are organized into more global concepts of face and facial features using the information of face geometry in order to reduce the ambiguity of features and determine the location of the face and facial features [18].

2.1.1.1.1 **Edges**

In computer vision, edge representation was applied in an early face detection system by Sakai et al. [19]; their work analyzed line drawings of faces from photographs, in order to locate facial features. Later, based on this work, Craw et al. [20] proposed a hierarchical framework to trace the human head line. Recent examples of edge-based techniques can be found in [21, 22]. For detecting edges, various types of edge detector operators are used and the Sobel operator is the most commonly used first derivative convolution filter, for detecting edges [22, 23].

A variety of second derivative filters (Laplacian of Gaussians) have also been used in some approaches in [24], while a large scale Laplacian was used to obtain lines in [25].

In a general face detector, which uses edge representation, labelling of the edges is needed. Then the labelled edges are matched to a face model. Govindaraju [26] accomplishes this

goal, for a front view of a face, by labelling edges as the left side, hairline or right side and then tries to match these edges against a face model by using the golden ratios of an ideal face.

2.1.1.1.2 **Colour**

In many face detection and related applications, human skin colour has been used and proven to be an effective feature. Using skin-colour as a feature for face detection has certain advantages; it is much faster than using other facial features. Skin colour is invariant to orientation so that face poses does not have too much influence. Skin colour differs among individuals, but several studies have suggested that the major difference exists in the intensity rather than the chrominance. Several colour spaces [27] have been used to label skin pixels including RGB, NRGB (normalized RGB), HSV (or HSI), YCrCb, CIE-XYZ and CIE-LUV colour spaces. Although, the effectiveness of the different colour spaces is arguable, the common point of all the above work is the removal of the intensity component. Terrilon et al. [25] recently presented a comparative study of several widely used colour spaces for face detection. Colour segmentation can basically be performed using appropriate skin colour thresholds where skin colour is modelled through histograms or charts [24, 28].

More complex methods make use of statistical measures that model face variation within a wide user spectrum. For instance, Oliver et al. [29] and Yang et al. [30] employ a Gaussian distribution to represent a skin colour cluster, consisting of thousands of skin colour samples, taken from the different human races. The Gaussian distribution is simply characterized by its mean and covariance matrix. Any pixel colour of an input image is compared with the skin colour model by computing the Mahalanobis distance. This

distance measure gives an idea of how close the pixel colour resembles the skin colour of the model.

2.1.1.2 **Template Matching**

Usually the standard typical human face is frontal in the template matching. Given an input image, the correlation values with several sizes of the standard pattern such as face contour, eyes, nose and mouth are calculated independently. These values determine the existence of the face portions. Although this approach has simplicity, it has however proved to be inadequate for face detection since it cannot handle variations in scale, rotation, pose and shape.

Sakai. et al. [31] reported an early attempt to detect frontal face from photographs , they use several sub-templates which correspond to facial feature such head line, eyes, mouth in order to model the human face.

The face template consists of six facial components consisting of two eyebrows, two eyes, a nose and a mouth. Face candidates are located by matching templates of face models represented by edges. In the final step, some heuristics are used to determine existence of a face. Experiments show better detection performance for images containing a single face, rather than multiple faces.

Lanitis et al. [32] established a detection method utilizing both shape and intensity information. In this approach, training images are formed in which contours are manually labelled by sampled points and vector sample points are used as shape feature vectors to be detected. They use a point distribution model (PDM) together with principal components analysis (PCA) to characterize the shape vectors over an ensemble of individuals. A face shape PDM can be used to detect faces in test images using an active shape model search to

estimate face location and shape parameters. The shape patch is then deformed to the average shape and intensity parameters are extracted. Then the shape and intensity parameters are used together in feature vectors compared using Euclidian distance.

2.1.1.3 **Generalized Knowledge Rules**

In generalized knowledge-based approaches, algorithms are developed based on heuristics about face appearance. Although it is simple to create heuristics for describing the face, the major difficulty is in translating these heuristics into classification rules in an efficient way. If these rules are over detailed, they may result in missed detections; on the other hand, if they are more general they may introduce false detections. In spite of this, some heuristics can be used at an acceptable rate in frontal faces with uncluttered backgrounds.

Yang and Huang [33] used a hierarchical knowledge-based method to detect faces. Their system consists of three level rules going from general to more detailed. Although this method does not report a high detection rate, their ideas for mosaicing (multi-resolution), and multiple level rules have been used by more recent methods.

2.1.1.4 **Appearance-based approaches**

A face in appearance-based methods is considered to be a single unit and features are extracted from the entire face region. An image feature (usually intensity) is extracted to construct a random variable x, which is a high dimensional vector in high dimensional feature space. Given the fact that face patterns are clustered in this high dimensional feature space and the probabilistic distribution of face pattern is unknown, the assumptions of Gaussian distribution or mixed Gaussian distribution leads to Bayesian classification or Maximum Likelihood (ML) approaches. Learning face appearances from samples has

attracted much attention recently and has provided excellent results, depending on the different statistical and neural models used.

In general, image-based approaches rely on machine learning and statistical analysis. Face detection is a two class (face, non-face) classification problem which relies on learned characteristics generally in the form of distributions.

The specific need for face knowledge is avoided by formulating the problem as a learning paradigm to discriminate a face pattern from a non-face pattern. Image-based approaches can be better understood by considering statistical supervised pattern recognition. A raw image can be taken as random variable x and this random variable is characterized by class-conditional density functions p(x|face) and p(x|non-face). If the dimensionality of x was not so high, a Bayesian or maximum likelihood classification would be possible. Hence, image-based approaches utilize more complex techniques such as subspace representations and learning networks to overcome the high dimensionality of the problem space.

Most of the image-based approaches apply a window scanning technique for detecting faces. The window scanning algorithm employs an exhaustive search of the input image for possible face locations at all scales, but there are variations in the implementation of this algorithm for almost all the image based systems. Typically, the size of the scanning window, the sub-sampling rate, the step size, and the number of iterations all vary depending on the method proposed and the need for a computationally efficient system.

To this end. Most of the approaches of the face detection and recognition aforementioned interpret the image candidate in term of intensity values [33, 34] and other transformed features such as edge, DCT coefficients , and wavelet coefficients.

Effective face detection involves identifying the intrinsic features that discriminate human faces from other objects in the feature space. In general the selections of the feature that represent the image content reflect the performance of the face detection algorithms. One of the major factors that affect the performance is the choice of suitable visual representation for facial features, the representation should be robust and easy to obtain.

Turk and Pentland[35], Sung and Paggio[36]. Rowley, et.al[34] described the face in terms of pixel intensity values. Some pre-processing steps were taken to compensate the variation caused the illumination condition, while Yullie et.al[37] interpreted the salient facial features using the peak and valleys in the image intensity and edge information. Papageorgion et.al[38], employed a set of wavelet coefficients corresponding to different scales to represent the human face. Podilchuk and Zhang [39], Wang and Chang[40] , Luo and Eleftheriadie[20] used DCT coefficient in the compressed domain.

In contrast to most geometric features, skin colour is insensitive to face orientation, occlusion, and pose. Due to these advantage skin colour is widely used to identify the face region[40, 41]. However skin colour features alone is not reliable to detect faces, and thus skin colour is used as pre-processing to asses filtering out non facial regions. and also can be combined with other features to detect faces.

Turk and Pentland,[35] projected a test image pattern onto a feature space the captured the variation learnt from the face sample. Measure distance was used between the test image pattern and the known faces in the feature space to identify faces. Sung and Poggio[36] frontal faces in complex background using view based face model which encodes face pattern using face and non face clusters. A neural network was used to classify by

evaluating the distances between a candidate region with the pre-trained face and non face clusters.

Multiple neural networks employed by Rowley et al [42] to search the face region in multiple positions and scale.They employs multiple networks system ; a router network first processes each input window to determine its orientation and then uses this information to prepare the window for one or more detector networks.

With the adoption of image and video compression such as JPEG and MPEG, several method have proposed to work on face detection and face recognition directly in the compressed domain

2.1.2 **Face detection in the DCT domain**

Performing analysis in the compressed domain (JPEG/MPEG) reduces the amount of effort required for decompression. Moreover, discrete cosine transform (DCT) coefficients, used as features, are attractive for pattern recognition since DCT based domain reduces spatial redundancy and gives compact information about patterns. Wang and Chang [40] combined skin colour, DCT coefficients which represent texture information to achieve high-speed face-detection without full decoding of the compressed video sequence.

To facilitate texture based detection , they converted the spatial domain algorithm proposed by Sun and Poggio[36] directly into DCT domain. Luo and Eleftheriadis [20] performed face detection using Sung's Gaussian mixture model [23] in the compressed domain. Chua, Zhao and Kankanhalli [43] proposed a face detection method that uses the gradient energy representation extracted directly from the compressed MPEG video data. To tackle the face recognition problem, Wang [40] used the input of the inverse quantized DCT coefficients of MPEG and generate bounding rectangle of the detected face region.

Among these approaches for face detection in the compressed domain, Luo's method [44] applying Sung's Gaussian mixture distribution-based face model [23] in the compressed domain for face detection achieved the best results. They treated face detection as a 1-dimensional vector classification problem.

In the DCT domain, feature vectors are created directly from (block based) DCT parameters. They use DCT values instead of image grey values as data of high dimension vectors to create the clustered Gaussian distributions, which model the face and non-face distribution in high dimensional space. After that, the system calculates the distances between input test image (getting vector from DCT values) and the centre of clustered Gaussian distributions and decides whether it is face or not from those distances.

However, a major problem to overcome in the compressed domain is that the image frames are divided into 8 by 8 blocks before DCT transform. Therefore, any detection work based on DCT parameters has to be done at the locations of blocks rather than pixels. That is the blocks reduce the spatial resolution of the system by 8, which makes it hard to detect small faces without fully decoding the image from the block based DCT parameters to pixels. This problem is called the block quantization (alignment) problem. In order to solve this problem, Luo [20] uses 16 out of 64 possible spatial alignment positions for each training face to generate extra training samples. However, this method gives extra variations, which do not belong to the face variations, to the face class and consequently gives a low detection rate.

From the review, many successful pixel domain face detection methods can be easily adapted to work in the compressed domain by using DCT coefficients as features. Even Gaussian distributions can be built on the DCT coefficients since the DCT is an

orthonormal transform and Euclidean distance and Mahalanobis distance can still be employed after the transform [20]. We also notice that most of current face detection methods are designed for grey images, not for the compressed domain. There are a few face detection approaches available for the compressed domain. However, they are not robust enough for the purposes of indexing. Therefore, developing a DCT domain face detection algorithm could be a possible direction for research. As mentioned above most successful pixel domain face detection methods can be easily adapted to work in the compressed domain by using DCT coefficients as features, by generating DCT coefficients features of Cb and Cr components directly (instead of grey intensity in pixel domain). In order to achieve this, a DCT coefficients feature extraction scheme is proposed. DCT coefficients, as features, are attractive for pattern recognition since DCT based compression reduces spatial redundancy and provides compact information about patterns. Moreover, it would be efficient if face detection can be implemented entirely in the compressed domain, without performing the inverse DCT followed by feature extraction, for thousands of compressed images. The features used for this purpose are the DCT coefficients of Cb and Cr components (chrominance) available from the compressed data. Since DCT coefficients capture frame information concisely, use of DCT features reduces the complexity of the neural network used in the algorithm. In addition, it improves the computational efficiency

2.1.3 **Neural networks**

Neural networks are powerful tools for solving complex non-linear classification problems and have been applied in many pattern recognition problems like object recognition face detection, where face detection can be referred as to class pattern recognition problem.

An early method of using hierarchical neural networks was introduced by Agui et al. [45]. The structure of the first stage consists of two parallel sub-networks in which the inputs are intensity values from an original image and intensity values from an image filtered using a 3×3 Sobel filter. The inputs to the second stage network consist of the outputs from the sub-networks and extracted feature values such as the standard deviation of the pixel values in the input pattern, the ratio of the number of white pixels to the total number of binarized pixels in a window, and geometric moments. An output value at the second stage indicates the presence of a face in the input region. Among all the face detection methods that have used neural networks, the most successful system image based face detection using neural networks was introduced by Rowley et al. [34] using skin colour segmentation to test an image and classify each DCT based feature vector for the presence of either a face or non face .

The neural network used in this thesis, the cascaded back-propagation neural network, was chosen because of simplicity and its capability for supervised pattern matching.

The structure of the neural network with three layers is shown in Figure 2.1.

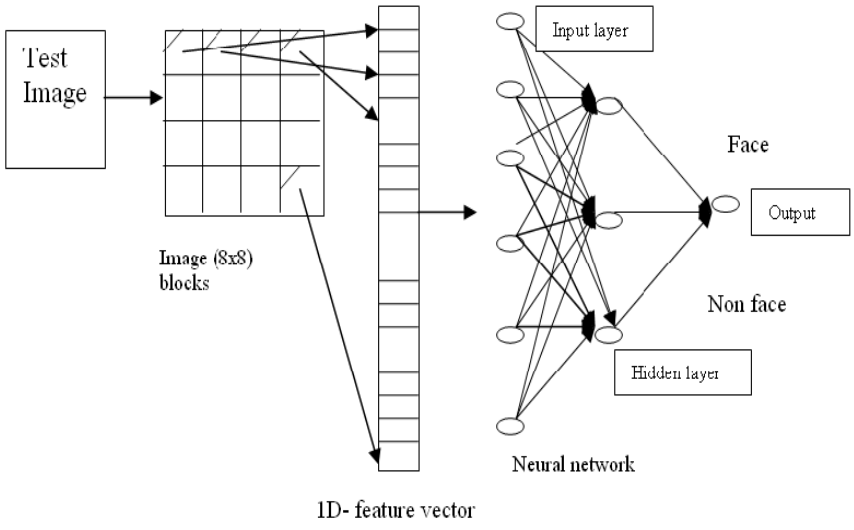


Figure 2-1 Neural network system

The input layer is a vector of 1x*N* DCT coefficient features to input nodes from each image either face or non face image, the hidden layers has *N* neurons and the output layer is a single neuron which is 0.9 if a face is presented and 0.1 otherwise. The classifier training process consists of supervised training. The extracted features represent the patterns and the desired outputs for each pattern are showed to the classifier sequentially. It processes the input pattern and produces an output. If the output is not equal to the desired one, the internal weights that contributed negatively to the output are changed by the back propagation learning rule; it is based on a partial derivatives equation where each weight is changed proportionally to its weight in the final output. In this way, the classifier can adapt its neural connections to improve its accuracy from the initial state (random weights) to a final state. In this final state the classifier should be able to produce correct or almost correct outputs. The network performance is measured by the Mean Squared Error (MSE). MSE is the sum of the squared absolute values of the difference between network outputs and desired outputs.

$$MSE_t = \frac{1}{patterns} \sum_{k=1}^{patterns} \left(\sum_{i=1}^{neurons} (n_{ki} - d_{ki})^2 \right) \qquad (2.1)$$

k is the pattern number and goes from 1 to the number of patterns; *i* is the number of the output neuron; *n* is the computed output and *d* is the desired output.

The desired outputs for the patterns are:

1) Skin face-like pattern: (0.9 0.1)

2) Non skin face-like pattern: (0.1 0.9)

In the presented work the neural networks is trained by feeding DCT coefficient feature vectors after skin face colour candidate obtained from the segmentation stage, which are the DC and the first three zigzag order AC coefficient feature samples from each block of 8x8 pixels of both Cb and Cr to classify each feature vector as output value 0.9 for a face and 0.1 for non-face.

2.1.3.1 **Cascade Correlation Neural Network**

The training of back-propagation neural networks is considered to be a slow process because of the step-size and moving target problems [46]. To overcome these problems cascade neural networks were developed. These are "self organizing" networks [46] with topologies which are not fixed and the grow of hidden layers during training. The supervised training begins with a minimal network topology and new hidden nodes are incrementally added to create a multi-layer construction. The new hidden nodes are added to make the most of the correlation between the new node's output and the remaining error signal that the system is being adjusted to eliminate. The weights of a new hidden node is fixed and not changed later, hence making it a permanent feature detector in the network. This feature detector can then be used to generate outputs or to create other more complex feature detectors [46]. In a CCNN, the number of input nodes is determined by the input features, while the number of output nodes is determined by the number of different output classes. The training of a CCNN starts with no hidden nodes. The direct input-output connections are trained using the entire training set with the aid of the back propagation learning algorithm. Hidden nodes are then added gradually and every new node is connected to every input node and to every pre-existing hidden node.

Training is carried out using the training vector and the weights of the new hidden nodes are adjusted after each pass [46]. Cascade correlation networks have a number of attractive features including a very fast training time, often a hundred times faster than a perceptron network [46]. This makes cascade correlation networks suitable for use with large training sets.

Depending on the application and number of input nodes, cascade correlation networks are fairly small, often having fewer than a dozen neurons in the hidden layer [47-49]. This can be contrasted with probabilistic neural networks which require a hidden-layer neuron for each training case. Also, the training of CCNNs is quite robust, and good results can usually be obtained with little or no adjustment of parameters [46].

2.1.4 **K-nearest neighbour**

The *k*-nearest neighbour algorithm (K-NN) is a machine learning algorithm for classifying patterns based on the closest training examples in the feature set. K-NN is a fast supervised machine learning algorithm which is used to classify the unlabeled testing set with a labelled training set [50]. The K-NN algorithm finds the minimum Euclidean distance from the training set samples. For example, given a query instance for an image; the K-nearest instances to this query image form the most common class. After finding the *k* nearest neighbours, the majority of these *k* nearest neighbours is taken to be the class output of the query image instance. The Euclidean distance D between two feature vectors X and Y is:

$$D = \sqrt{\sum_{i=1}^{N}(X_i - Y_i)^2} \qquad (2.2)$$

where x_i and y_i are elements of X and Y, respectively.

In order to classify an image, the training phase of the K-NN algorithm consists of gathering the feature vectors and class labels of the training samples. In the testing phase, k is a user defined variable, the unlabelled vector of the query image is classified by assigning the label which is the most frequent among the *k* training samples nearest to that query image. The KNN algorithm procedure is summarized as follows:

1) Determine the value of k parameter (k = nearest neighbours)
2) Compute the Euclidean distance between the query image and all the training samples.
3) Sort the distances and determine the nearest neighbours based on the k^{th} minimum distance.
4) Gather the class of the nearest neighbours.
5) Find the output class of the query image by using the simple majority of the classes.

2.2 Verification and Validation Techniques

2.2.1 Jack-knife Technique

The Jack-knife technique [51] is usually implemented to provide a correct statistical evaluation of the performance of a classifier when applied to a limited number of samples divided into two sets: a training set and testing set. It was employed to evaluate the performances of the learning system used in this work. In practice, a random number generator is used to select the samples used for training and the samples kept for testing. The classification error varies with the training and testing sample sets and, for a finite number of samples, an error-counting procedure is used to estimate the performance of the classifier [51]. In this work, 80% of the available samples were randomly selected and used for training while the remaining 20% were used for testing. The results were then analyzed to assess the performance.

2.2.2 **Performance Criteria**

The values for negative and positive results in an experiment are likely to be the most useful practically. The performance criteria used in this work are accuracy (the fraction of all correct predictions), sensitivity (the fraction of positive cases correctly classified) and specificity (the fraction of negative cases correctly classified) [52]. These are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.3)$$

$$Specifity = \frac{TN}{TN+FP} \qquad (2.4)$$

$$Sensitivity = \frac{TP}{TP+FN} \qquad (2,5)$$

The numbers of true positive, true negative, false positive and false negative results are indicated by TP, TN, FP and FN respectively, which are defined as follows:

$$TP = \frac{correct\ positive\ predidtion}{Total\ positive} \qquad (2.6)$$

$$FP = \frac{incorrect\ \ negative\ \ predidtion}{Total\ \ negative} \qquad (2.7)$$

$$TN = \frac{correct\ \ negative\ \ predidtion}{Total\ \ negative} \qquad (2,8)$$

$$FN = \frac{incorrect\ \ positive\ \ predidtion}{Total\ \ positive} \quad (2,9)$$

In these equations the numerators and denominator terms are defined as follows: “Correct positive predictions” is the total number of cases for which the system makes correct predictions; “Incorrect positive predictions” is the total number of cases for which the system makes an incorrect predicted; “Correct negative predictions” is the total number of cases for which the system correctly predicts a non match; “Incorrect negative predictions” is the total number of cases for which the system incorrectly predicts a non-match; “Total positives” is the sum of correct matching cases (Number of associated cases used in testing); “Total negatives” is the sum of correct non-matching cases (Number of un-associated cases used in testing) [52, 53].

2.3 Survey of compressed based DCT domain

2.3.1 DCT based JPEG Compression Standard

JPEG (Joint Picture Expert Group) is a very well known ISO/ITU-T standard created in the early 1990s. There are several modes that are defined by the JPEG standard [54], but all JPEG file readers must be capable of reading the baseline mode. As the image files are stored in JPEG extension, the image is divided into non overlapping blocks (8x8) pixels and each of these is transformed using the Discrete Cosine Transform (DCT) into a set of 64 DCT coefficients. The DCT, which was established in [55], plays an extremely important role in image and video coding. The zero coefficient, $C_{(0,0)}$, is called the DC coefficient, while the other 63 coefficients are referred to as AC coefficients. The pixels energy tends to be packed into the upper left (low frequencies) of the array of DCT coefficients. Each of the resulting DCT coefficients is uniformly quantized according to the corresponding values from the quantization table.

In the compressed domain, the DCT coefficients as features have the following attributes:

- The DCT is an orthonormal transform; both the Euclidean distance and the Mahalanobis distance are unchanged after the transform, only if all frequencies are used. Therefore, statistical models or other pattern recognition methods may not be influenced by converting the problem to the DCT domain
- The DCT domain is more attractive than the pixel domain for pattern classification problems because the DCT transform reduces the correlation among individual components and compresses the feature energy in the low frequency parameters. In other words, DCT coefficients contain compact information about the pattern of interest. Moreover, it is much

easier to choose feature components directly from DCT parameters rather than from pixel values from the compressed image. Therefore it is practical to directly use DCT coefficients as features to classify patterns (human faces).

The 2D DCT transform can be defined as follows:

$$C(u,v) = \alpha(u)\alpha(v)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} x(i,j) \quad cos\left(\frac{(2i+1)u\pi}{2N}\right) cos\left(\frac{(2j+1)v\pi}{2N}\right) \qquad (2.10)$$

Here *i* and *j* are spatial coordinates of the image *x* (*i*, *j*), *u*, *v* are the coordinates of the DCT coefficients C(*u*,*v*) and α is defined as

$$\alpha(u) = \alpha(v) = \begin{cases} \frac{1}{\sqrt{N}}, & for\ u = 0 \\ \frac{2}{\sqrt{N}}, & otherwise \end{cases}$$

The corresponding inverse transform is defined as

$$x(i,j) = \sum_{u=0}^{N-1}\sum_{v=0}^{N-1} \alpha(u)\alpha(v)c(u,v) \quad cos\left(\frac{(2i+1)u\pi}{2N}\right) cos\left(\frac{(2j+1)v\pi}{2N}\right) \qquad (2.11)$$

Each block is DCT-transformed. The DC and the other AC coefficients are picked in zigzag order as shown in Figure 2.2. These coefficients are characterised by: (1) the DC coefficient of each sub-block represents its average and hence describes its energy which is an important property in image processing [56]; (2) the remaining AC coefficients contain frequency information which describes different patterns of content.

Extracting features is an attempt to reduce the high dimensional image data space into fewer dimensions, which is essential for effective features. The most popular techniques used are principle component analysis (PCA) and linear discriminate analysis (LDA); however these methods have high computational cost and require large numbers of training

samples [57]. The DCT transform domain is a way of resolving the problem of storage space and the cost of computational complexity.

2.3.1.1 **Coefficient quantization**

In general, each of the 64 DCT coefficients is uniformly quantized in conjunction with a 64-element Quantization Table, which must be specified by the application (or user) as an input to the encoder. Each element can be any integer value from 1 to 255, which specifies the step size of the quantizer for its corresponding DCT coefficient. The purpose of quantization is to achieve further compression by representing DCT coefficients with no greater precision than is necessary to achieve the desired image quality. Stated another way, the goal of this processing step is to discard information which is not visually significant. Quantization is a many-to-one mapping and therefore is fundamentally lossy. It is the principal source of data loss in DCT-based encoders. Quantization is defined as division of each DCT coefficient by its corresponding quantization step size, followed by rounding to the nearest integer.

$$I^{Q}(u,v) = integerRound\left(\frac{I(u,v)}{Q(u,v)}\right) \qquad (2.12)$$

The JPEG compression standard requires uniform quantizers to be used for all the DCT coefficients. A matrix is used to specify the quantization of the DCT coefficients, where the entry, *Q(u, v),* in the matrix gives the quantization step size for the DCT coefficient *I(u, v).* Quantization of the DCT coefficients achieves image compression, but also causes distortion in the decompressed image. Specifically, quantization of coefficient induces an error image which is simply the associated basis function, with amplitude equal to the coefficient quantization error.

2.3.1.2 **Zigzag Order**

All the transformed coefficients output from the DCT process, the 64 DCT coefficients, are ordered into the "zig-zag" sequence as shown in figure 2.2. This ordering helps by placing low-frequency non-zero coefficients before high-frequency coefficients. The DC coefficient, which located on the upper left, contains a significant fraction of the total image energy. An effective scheme, called the zigzag scan ordering, was proposed by [58] which converts the 2-D array of transform coefficients into a 1-D sequence. Applying zig-zag scanning, all successive zero high frequency coefficients tend to come together, which makes it easy to apply run length coding and by doing this, the coefficients are arranged in low, mid and high frequency order.

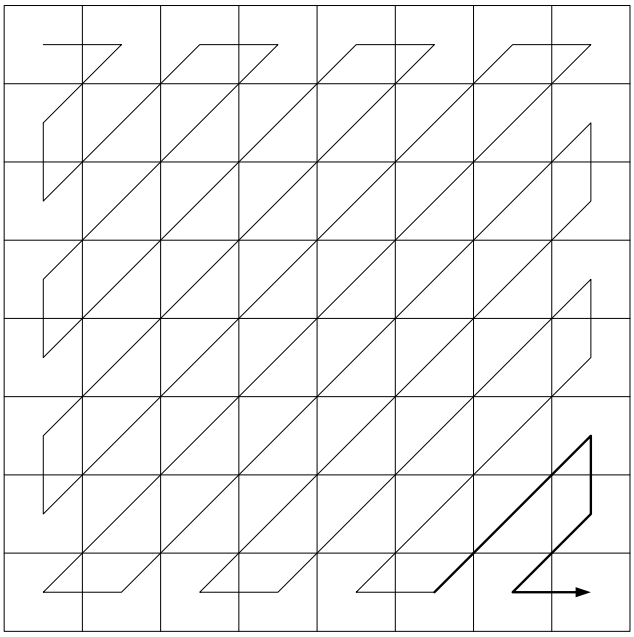


Figure 2-2 DCT Zigzag order

2.3.2 **The Discrete Wavelet Features**

In signal analysis and image processing, the discrete wavelet transform (DWT) is a very useful tool in many applications [59]. The DWT is a hierarchical sub-band technique, which is widely used in multi resolution pattern recognition [60] and is employed to extract features from an image. In the frequency domain, the signal is decomposed into sinusoidal components. The one-dimensional discrete wavelet transform (1-D DWT) decomposes an input sequence into approximation and detail sub-bands by calculations with a low-pass filter and a high-pass filter respectively. The two-dimensional discrete wavelet transform (2-D DWT) decomposes an input image into four sub-bands, one average component (LL) and three detail components (LH, HL, HH) as shown in Figure 2.3. These sub bands are labelled as follows:

- Sub band LL represents the horizontal and vertical low frequency components of the image which are known as approximation coefficients.
- Sub band LH represents the horizontal low and vertical high frequency components of the image which are known as vertical coefficients.
- Sub band HL represents the horizontal high and vertical low frequency components of the image which are known as horizontal coefficients.
- Sub band HH represents the horizontal and vertical high frequency components of the image which are known as diagonal coefficients.

Figure 2-3 The result of 2-D DWT decomposition.

More information about the 1D DWT can be found in [59]. A decomposition of 2D DWT is computed with a separable extension of the 1D decomposition. The basic steps for the 2D DWT decomposition is shown in Figure 2.4 [61].

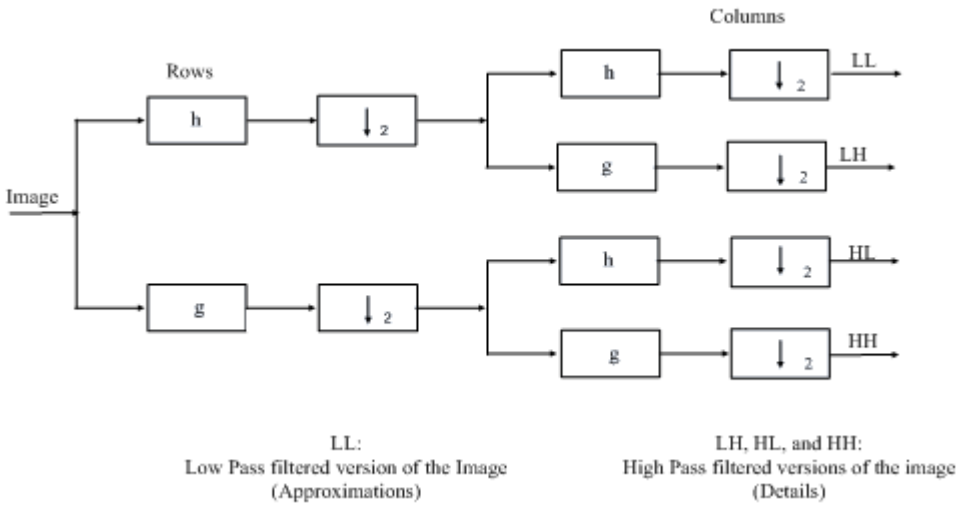


Figure 2-4 A wavelet decomposition of an image.

A low-pass filter and related high-pass filter are applied to each row and column of the input image. Each filtered row/column is sub-sampled by a factor of 2, throwing away half the filtered data. The two types of output, the low-pass samples and the high-pass samples, are grouped separately. The mechanism can be repeated on the low-pass filtered samples providing data corresponding to a lower resolution.

In this research, the 2D DWT is used to extract the features, as it is well known that DWT coefficients can provide a powerful insight into an image's frequency and spatial characteristics. Given an image $f(i, j)$, its 2D DWT transform is defined as follows:

$$f(u, v,..) = \sum_{i,j} f(i, j) g_{u,v,...}(i, j) \qquad (2, 13)$$

where i and j are the spatial variables and u, v,.. are transform domain variables.

The coefficients of the 2-D DWT are the wavelet features. Usually, an input image is decomposed by i^{th} levels of wavelet transform. At the i^{th} level, each image can be decomposed into four sub-bands namely LL_i , LH_i , HL_i and HH_i . The LL_i sub band represents the low frequency components, and HH_i represents the high frequency components. Figure 2.5 illustrates the 2D DWT decomposition of an image decomposed by 3 levels of wavelet transform.

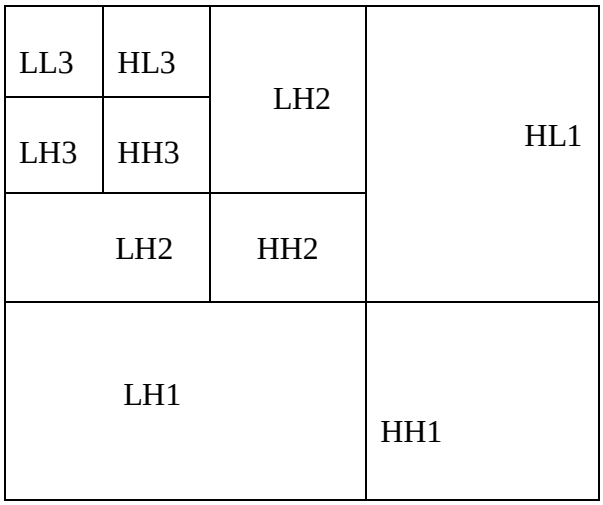


Figure 2-5 2D DWT decomposition of an image

2.4 Survey of Content Based Image Retrieval

2.4.1 Content Based Image Retrieval

Nowadays, the new technologies providing access to the internet and external storage capacity have become relatively cheap, making it more and more economic to share online and store huge image collections, such as satellite images, medical images, family albums and general collections of images. Furthermore online photo sharing has become extremely popular [62], involving hundreds of millions of pictures with different content. Also video sharing using YouTube has brought a new wave of content based image multimedia information retrieval. In the past decade a number of commercial content based image retrieval systems have been developed, such as QBIC [1], Photobook [3], virage [63], visualSEEK [4], SIMPLIcity [5], Blobworld [64], etc. However, in these systems, the retrieved images from the image database rely on one query image, while the images that the user request are often similar to a set of images with the same conception. In addition, most of the existing approaches assume that there is a linear relationship between different features, and the effectiveness of such systems is limited due to the difficulty in representing high-level concepts using low-level features.

The retrieval system is located between the query information requested by the user and the information content of the database from which images are to be retrieved [65]. The modules of the system are composed of: Querying, which enables the user to make an inquiry; Description, for capturing the essential content/meaning of the request and transferring it to an internal representation; Matching, the metric measurement to enable searching for the required image information in the database; Retrieval, for retrieving the

relevant images from the database; and Verification, for ensuring/confirming the relevance of the retrieval results.

There are four kinds of distance measurement [66], depending on the methods that are used to compute the distance between the query image and the image in the database images, including colour similarity, texture similarity, shape similarity and relationship similarity.

Content based image retrieval has a broad scope for application, and various procedures are used. Intone [67] used a list of keywords as queries, since in his database each image is annotated by some keywords (annotation words) which explain the content of the image, and this is called query-by-text (QBT). QBT is practical when users want to freely express their search needs. In other applications, queries may be issued by providing example images, and this type of image retrieval is therefore called query-by example (QBE). Most QBE retrieval is "content-based image retrieval" [68].

Content-based image retrieval is the retrieval of relevant images from an image database based on features automatically derived. The need for this type of approach has increased enormously in many applications including commerce, biomedicine, crime prevention, military, culture, education, entertainment, and web image classification and searching. For this reason, content-based image retrieval has been widely studied. However, space limitations do not allow this thesis to present a complete survey. Instead, emphasis is placed on some of the literature that is most related to the work that I propose.

Due the large sizes of significant image collections, and the difficulties faced by manual annotation in text based image retrieval which is highly labour-intensive, time consuming, and impractical with large databases, content based image retrieval was proposed early in 1990 to overcome these difficulties. That is, instead of being manually annotated by text

key words, images would be represented or indexed by their own visual content, such as colour, texture, and shape. Since that time, many techniques and many image retrieval systems have been developed following this research direction.

A typical Content-based Image Retrieval (CBIR) system not only deals with various sources of information in different formats (for example, text, image, video) but also user's requirements. Basically it analyzes both the contents of the source of information as well as the user queries, and then based on suitable metric measurement matches these to retrieve those items that are relevant.

The major functions of such a system of CBIR are the following:

- Analyze the contents of the source information, and represent the contents of the analyzed sources in a way that will be suitable for matching user queries (space of source information is transformed into feature space for the sake of fast matching in a later step). This step is normally very time consuming since it has to process sequentially all the source information (images) in the database. However, it has to be done only once and can be done off-line.
- Analyze user queries and represent them in a form that will be suitable for matching with the source database. Part of this step is similar to the previous step, but applied only to the query image.
- Define a strategy to match the search queries with the information in the stored database. Retrieve the information that is relevant in an efficient way.

2.4.1.1 **Feature extraction**

Image features (content) are the basis of content based image retrieval (CBIR), in a broad sense, features may include both text-based features (key words, annotations) and visual features (colour, texture, shape). Another definition of a feature is to capture a certain visual property of an image, either globally for the entire image or locally for a small group of pixels. The most commonly used features include those reflecting color, texture, shape, and salient points in an image, Furthermore, the features can be classified as general features and domain specific features. The former include colour, texture, and shape features while the latter is application-dependent and may include, for example, human faces and finger prints. According to [69], due to the variety of image content and the subjectivity of various applications, there exists neither a universal feature valid for all images nor a specific best representation for a given feature. In their work they concluded that the retrieval performance of features is dependent on the application whereby features like colour histograms and invariant feature histograms are suitable for a database of arbitrary colour pictures; the best feature may change from image to image and from application to application.

Image processing research has gone through many years of development in areas such as enhancement, segmentation, feature extraction and pattern classification. Moreover the image database is often compressed due to the limitation of the storage size and the network bandwidth. The Joint Photographic Experts Group (JPEG) is widely used on the World Wide Web, because of its good compression rate and image quality. To bridge the gap between the compressed domain and the pixel domain, where the majority of image processing has been developed, recent research has started to develop image analysis and

content feature extraction algorithms directly from the DCT based compressed domain [7, 70, 71].

2.4.1.1.1 Colour

Colour is the most widely image feature in content based image retrieval, which is relatively robust to background and invariant to image size and orientation. Colour can be indexed by feature descriptors such as colour moments [72] and colour histograms [73]. The colour histogram is most commonly used for feature image representation. The colour histogram is obtained by summing the number of colour coefficient with similar values in the colour space component. Furthermore, the most common colour spaces are RGB, CIE Lab, CIE Luv, HSV and YCbCr.

2.4.1.1.2 Texture

Texture is another feature that has been widely used in CBIR; it refers to the innate surface properties of an object and their relationship to the surrounding environment. Examples of common textures include clouds, trees, bricks, hair, etc. [74]. In the early 1970s, Haralick et al. proposed the co-occurrence matrix representation of texture features [75]. Tamurra texture [76, 77] was subsequently proposed as an enhanced version. In the early 1990s, after the wavelet transform was introduced, researchers began to study the use of wavelet transforms in texture representation [77, 78]. Among retrieval methods a popular way to form texture features is by using the coefficients of a certain transform on the original pixel values, or, more sophisticatedly, by statistics computed from these coefficients. Examples of texture features using the wavelet transform and the discrete cosine transform [78].

2.4.1.1.3 **Shape**

Shape is a feature that represents the contour of an object in an image. It is invariant to the size and location of the object. Shape feature extraction involves the segmentation of images into regions or objects. However, it is difficult to achieve accurate image segmentation and to extract the contour of an object correctly. Shape can be represented by feature descriptors such as Fourier descriptors, chain codes, moment invariants and Zernike moments [79].

2.4.1.2 **Image Retrieval Evaluation**

A number of investigators have highlighted the advantages offered by the use of user-centred evaluation techniques in image and information retrieval. The evaluation of image retrieval involves determining the appropriate dataset to be evaluated, which should be large enough from the user point of view. In addition, the system should be evaluated by an appropriate metric measure in order to try to model the human requirement's perspective. Standard precision and recall metrics merely show the retrieval effectiveness of the underlying system and do not take account of user interface and speed issues. Therefore, if such measures were used to compare two image retrieval systems, they would not be able to predict which system would perform best under certain task conditions and with different groups of users. User centred evaluation however, allows us to compare system performances with different users and for a large variety of tasks [73]. This view was also supported by [80], who stressed the need for evaluating real world systems in real world settings.

2.4.1.2.1 **Precision and recall**

It is ideal to retrieve all the relevant images in a large image database to evaluate the effectiveness of image retrieval. However, it will be very time consuming and is impractical. In many experiments therefore, the evaluations are calculated based on a few images retrieved during a query by calculating the well known precision and recall which are widely used as statistical measurement classifications.

$$precision = \frac{C}{B} \qquad (2, 14)$$

$$recall = \frac{C}{A} \qquad (2.15)$$

Here *A* is the total number of relevant images in the collection, *B* is the number of images retrieved and *C* is the number of correctly relevant or similar correctly images retrieved. These are standard measures in image retrieval, which give a good indication of system performance. However, neither precision value nor recall value alone contains sufficient information. We can always make recall 1, simply by retrieving all images. Similarly, precision can be kept high by retrieving only a few images. Thus precision and recall should either be used together or the number of images retrieved should be specified. Precision and recall are often averaged, but it is important to know the basis on which this is done [81].

2.4.1.3 **Receiver Operating Characteristic (ROC)**

ROC curves are graphical plots of the probability of correct detection (true positive rate TPR) against the probability of false detection (false positive rate FPR). ROC curve analysis is a method of comparing classifiers on natural datasets using accuracy [82]. The ROC graph gives significant indication when the image retrieval system process straightforward. The trade-off between True Positive (TPR) and False Positive (FPR) and

its derived ROC curve is plotted using the TPR and FPR values from the testing dataset for each given acceptance threshold value. The acceptance threshold values are computed using the images in the training dataset.

2.4.1.4 **DCT coefficient based image retrieval**

Most images in databases are distributed or stored in JPEG format at the source. It can significantly reduce computational effort if image retrieval is performed in the DCT domain avoiding the full data decompression process.

Research in the DCT domain image or video processing has become more and more important. Recent research has developed image analysis and content features extraction algorithms directly from the DCT domain [7], [70, 71, 83] and wavelet domain [83].

The DCT coefficient values can be regarded as the relative amounts of the 2D spatial frequencies contained in the 8x8 block of input data Therefore, given 8×8 blocks from a compressed image, we need only use the low spatial frequency coefficients to construct an energy histogram. Using histograms is a popular analysis method used in the image retrieval field, which was introduced, first by Swain and Ballard [73] via the use of colour histograms.

The DCT energy histogram is constructed by counting the number of times an energy level occurs in (8×8) DCT blocks. After constructing the DCT energy histogram, the histogram intersection metric method can be used to match the reference image DCT energy histogram to those of different groups in the databases. The selected images will have higher matching values. Histogram intersection is an efficient way to perform matching. Its computational complexity is low and it can be implemented on most computers. The matching similarity algorithm proceeds as follows. Given a pair of histograms, h and g,

each containing DCT energy levels divided into *n* bins, the intersection of the histograms, (proposed by Swain and Ballet [73]), is defined as:

$$\sum_{j=1}^{n} \min(h(j) - g(j)) \qquad (2,16)$$

The matching result is normalized to the range 0 to 1, by dividing the total number of coefficients used in the reference image as follows:

$$H(h,g) = \frac{\sum_{j=1}^{n} \min(h(j) - g(j))}{\sum_{j=1}^{n} g(j)} \qquad (2,17)$$

where $h_{image}(j)$ and $g_{query}(j)$ *are* the j^{th} feature and n represents the number of features. Observe that the larger the similarity distances the better the fit.

2.4.1.5 **DCT characteristic**

Among all the compressed formats the DCT is the most widely adopted (JPEG/MPEG, H261/H263) [71, 84] because: (a) The DCT is close to the optimal Karhunen Loeve transform (KLT), (b) The DCT is signal independent and has the ability to overcome the weakness of the KLT (C) DCT fast algorithms are readily available for efficient implementation [85].

Following application of the DCT operation in a given input image, each block produces a set of 64 signal amplitudes (in terms of 64 basis signals), which make up the signal's 2D frequency spectrum. Since the DCT transform is a one-to-one mapping, the overall size of the DCT result obtained is equal in size and in dimension to the input image, but can be

viewed as a block by block signal representation of the original image, with a number of interesting properties [54] as follows:

1. The first coefficient of a given block (DC) represents the average pixel intensity of that block.
2. Low frequency coefficients are located in the upper left of the DCT coefficients.
3. High frequency coefficients appear later using the zigzag ordering (bottom lower right) of the DCT coefficients.
4. Typical 8 x 8 blocks do not vary greatly over such small distances, therefore the higher frequency contributions (amplitudes appearing later in the coefficient sets) tend to be negligible and are particularly well suited for compression.

2.5 Survey of semantic based image retrieval

2.5.1 Semantic based image retrieval

Many algorithms have been designed to describe colour, texture, and shape features. These algorithms cannot sufficiently model image semantics and have many limitation when dealing with broad content image databases [86]. Extensive research experiments on CBIR systems show that low level contents often fail to describe the high level semantic concepts in user perception.

In order to improve the retrieval accuracy of content-based image retrieval systems, research focus has been shifted from designing sophisticated low-level feature extraction algorithms to reducing the 'semantic gap' between the visual features and the richness of human semantics [88].

To describe image semantics, Eakins in [89] classified queries into three levels of abstraction. Level 1 retrieval is applied using primitive features such as colour, texture, and shape. A typical example query is "find pictures like this". Essentially at this level no semantic information is used in an image and is related to CBIR. Level 2 is retrieval of objects identified by derived features which involve some degree of logical inference. This level includes object semantics such as object classes and spatial relations among objects. Those systems which can resolve this level of queries are considered as retrieval by semantic content. The third level comprises retrieval by abstract attributes, involving a high degree of abstract and possibly subjective reasoning about the meanings and purposes of the objects or scenes depicted. Examples in this level include scene semantics, behaviour semantics, and emotion semantics. Eakins's classification is helpful for describing the

capabilities and limitations of different retrieval techniques. Level 2 and 3 are considered as semantic image retrieval [87] and the difference between Level 1 and 2 as the ‘‘semantic gap’’.

The question is how to bridge the semantic gap between low level and high level features.

Liu et. al[87] reported that the state of art technique of linking the gap mainly including five categories: (1) using object ontology to define image concept (2) machine learning to associate low level features with query concept (3) Relevance feedback, introducing the user into the retrieval loop for continuous learning (4) generating semantic template to support high level image retrieval (5) integrate the visual content with the textual information.

In object ontology, semantics can be derived easily from our daily language. For instance, sky can be described in image as upper, uniform and blue region area. Different intervals can be defined for the low level features to the systems that using semantic, for example specific colour description interval such ‘light green, medium green , dark green’ all of these description derived from simple daily language can be referred to object ontology which is provide a qualitative definition of high level query concept. In Ref [90] presented atypical example of ontology system. In their system each region in the image is described by it average colour in lab colour space. vertical and horizontal axis positions, size and shape .

In machine learning, supervised learning and unsupervised learning are used in most cases to derive high level semantic features from the image content. Due to limit of the thesis space unsupervised learning is not our focus point. In supervised learning different classifier techniques such as support vector machine (SVM)[91], Bayesian networks

classifier[92], and neural networks are often used to learn high level concepts from low level features. In the training process need dataset of features either low level features or segmented regions that characterize the image content categories are fed into the neural networks classifier to establish classification in order to link low features of an image with the semantic category label that represent the high level feature. The problem of their system large a mount of training dataset is required and intensive computation.

A semantic template can be defined as a representative feature of a concept, where the concept refers to image contents such as object regions derived from a collection of sample images. Reference [93] introduced the semantic visual template in order to link the low level features with the high level features. A visual template is a set of examples of objects or scenes that represent specific image concepts such as sunset, horse, etc.

Relevance feedback (RF) was introduced to CBIR in the mid 1990s and it is powerful tool used in the text information retrieval in order to bring the user in the retrieval loop to reduce the semantic gap between the levels of query that represent the content image and what the user thinks. Relevance feedback provided significant performance by continues learning through interaction with the end user CBIR. Where the system of CBIR provide initial results using query by example sketch etc. the user then judge that the outcome results whether or what degree relevant to the query. Based on the user feedback the machine learning applied to learn the feedback. In this thesis the focus is on generating knowledge template and using the machine learning, particularly neural networks in order to narrow the semantic gap between the low level features and the high level features as will be a addressed in chapter 3, 4, 5.

CHAPTER Three

3 Methodology of content based image retrieval

3.1 An Efficient Feature Extraction Technique based on a histogram of Block DCT Coefficients

In recent years, along with the large demands for digital photography and multimedia systems, the sizes of image volumes and video collections have increased rapidly. As a result, the rich source of visual information made available in our daily life has brought the need to develop algorithms and techniques to browse, search, classify, retrieve and manage the content of those compressed digital images and videos.

Content based image retrieval (CBIR) has been a hot topic research over the last decade. Most digital images are stored in the compressed format defined by joint picture expert group (JPEG), widely used on the internet and in image databases[94]. Nowadays, the need to organize, search, classify and retrieve images has become an impo**r**tant issue. Traditional image retrieval techniques for digital images have used text-based systems to classify and retrieve images from databases; this kind of technique creates metadata about images, and this process, as a result, can be highly labour-intensive, time consuming, and produce results inconsistent with human perception. For content-based image retrieval, the main idea is to extract low level features, such as colour, texture, and shape, which are representing the image properties[94].

Performing image retrieval or classification in the spatial domain may not be as efficient as that directly performed in the DCT domain. To obtain fast feature extraction for compressed domain images, a new wave of research effort directly provides access to feature extraction in the transform domain[12, 95]. All existing research related to the

compressed domain is limited to DCT domain. The reason is that DCT can be considered as a good approximation of principal component extraction, which helps to process and highlight the signal frequency features. Recent research is started to develop image analysis and content feature extraction directly from compressed domain[9, 12, 96].

In CBIR, the colour visual properties are widely used as features, which are relatively more robust in the presence of complicated backgrounds than any other features and it is invariant to image rotation or orientation. Histogram quantization is referred to as a method of reducing the size of colour information. Various quantization methods are being studied to extract colour features; the most common one is the traditional colour histogram technique[73].

Vadivel et .al [97] combined two features colour and texture, colour histogram is extracted from image colour information, while Haar or Daubechies wavelets is extracted to represent the texture feature of the image. Another histogram based approach can be found in[98], where the so-called blob world is used to search similar images. Wang and Yong used energy to exploit correlation between wavelet packet sub-bands and derive discriminating features[99].

In this work, the proposed method is a DCT histogram-based feature extraction technique for image classification. Histogram quantization features are extracted directly from the DCT coefficients computed over image blocks. These coefficients represent the low frequency content of the input image. This method extracts and constructs a feature vector of histogram quantization from a few DCT coefficients in order to limit the number of coefficients having the same DCT coefficient over all image blocks, or in other words, constructing a histogram by counting the number of blocks with the same DCT coefficients.

The extracted features for each image block in any colour space are concatenated to construct a feature vector. The database image and query image are equally divided into non overlapping 8×8 blocks of pixel. Therefore, each of them is associated with a histogram quantization feature vector derived directly from the DCT coefficients. Users can select any query as the main theme of the query image. The retrieval of images is based on the similarity between a query image and any database image, which is ranked according to the closest similar measures computed by Euclidean distance.

3.1.1 **Proposed Feature Extraction Techniques**

The proposed method extracts a quantized histogram from DCT coefficients in the transform domain as a feature vectors to represent the content of the input image. The input image is first divided into 8×8 non overlapping blocks. Then, each block is DCT-transformed. The DC and the first three AC coefficients are picked in a Zigzag order. Not all the DCT coefficients contain useful information. Only the coefficients that are located in the upper left corner in each sub-block are considered. These coefficients are characterized by: (1) the DC coefficient of each sub-block, which represents its average and hence describes its energy; (2) the remaining AC coefficients contain frequency information which describes patterns of different content.

In the proposed technique, in order to obtain a better quantization of the coefficients that have been selected, normalization is applied by calculating the minimum and maximum values of the entire selected coefficients over the whole database. These are denoted by DC_{min} and DC_{max} the minimum and maximum values of the DC coefficients for the whole database and by L the number of quantization bins. The size of each bin b_{DC} is given by

$$b_{Dc} = \frac{(DC_{max} - DC_{min})}{L} \qquad (3.1)$$

Likewise, b_{Ac1}, b_{Ac2} and b_{Ac3} are expressed by

$$b_{Ac1} = \frac{(AC1_{max} - AC1_{min})}{L} \quad (3.2)$$

$$b_{AC2} = \frac{(AC2_{max} - AC2_{min})}{L} \quad (3.3)$$

$$b_{AC3} = \frac{(AC3_{max} - AC3_{min})}{L} \quad (3.4)$$

In the experiments L is set to 32 based on empirical assessment. Notice that this quantization makes each coefficient in a specified range.

The histogram DCT quantization feature vector is constructed for both query and database images by computing the number of DCT quantized coefficients belonging to each level from 1 to L.

$$H_{Dc} = \{prob_{Dc}(q_1), prob_{Dc}(q_2), \ldots\ldots\ldots\ldots\ldots\ldots, prob_{Dc}(q_L)\} \quad (3.5)$$

$$H_{Ac1} = \{prob_{Ac1}(q_1), prob_{Ac1}(q_2), \ldots\ldots\ldots\ldots\ldots\ldots, prob_{Ac1}(q_L)\} \quad (3.6)$$

$$H_{Ac2} = \{prob_{Ac2}(q_1), prob_{Ac2}(q_2), \ldots\ldots\ldots\ldots\ldots\ldots, prob_{Ac2}(q_L)\} \quad (3.7)$$

$$H_{Ac3} = \{prob_{Ac3}(q_1), prob_{Ac3}(q_2), \ldots\ldots\ldots\ldots\ldots\ldots, prob_{Ac3}(q_L)\} \quad (3.8)$$

where, $prob_x(q_i)$ represents the probability of level q_i, which corresponds to the coefficient of type X. All channel colour components of YCbCr feature vectors are concatenated to form the final feature vector of the input query image.

3.1.2 **Image Classification**

Experiments were performed on the Wang-database[100], which consists of ten classes, namely: African People, Beach, Building, Buses, Dinosaur, Elephant, Flower, Hours, Snow Mountain and Food. Each class contains of 60 images.

Let $C_1.C_2$…….C_k be the classes of the database images. In image classification, each query image should be assigned to a class membership label K which indicates the class it belongs to. Our classification is performed by comparing the histograms of specific DCT quantization coefficients of test query and database images. Large image database systems mostly require an efficient comparison measure as well as feature extraction in order to provide a reasonable response to an image query. After deciding on features to represent images, features from the query images need to be compared against all the image dataset features to find matches. The similarity measure for a given query image involves searching the database for similar histogram quantization feature vectors as the input query. Euclidean Distance is a suitable and effective method which is widely used in the image retrieval area. The retrieval results are a list of images ranked by their similarity distances to the query image. The similarity distance measure between the vectors of query image and the database image is defined as

$$D(I_q, I_d) = \frac{\sqrt{\sum_{i=1}^{N}(I_{q_i} - I_{d_i})^2}}{N} \qquad (3.9)$$

where D is the distance between the feature vector and, N represents the number of DCT blocks. The distances values are ranked in ascending order. Then, the KNN (K Nearest Neighbour) classifier is used to classify the query image. Classifying a query image to specific sample class (for instance, if the query image for any class image ranges for example from the lower limit and upper limit of the class) to which class it is closest in feature space of the image database based on the distance. By the k-nearest neighbour rule, for k = 10, the proposed system returns or classifies the nearest ten neighbours (those ten with the least computed distance between the query image samples and the reference

database image). Or in other words, the KNN classify images to the top number of images that belong to query image.

3.1.3 **Experimental Results**

To evaluate the classification efficiency of the proposed features, the DCT Histogram quantization features were compared with a traditional spatial-domain colour-histogram classification technique[73] and the modified colour histogram commonly known as the annular approach [101]. For the sake of demonstration, both RGB and YCbCr colour spaces have been used. It is worth mentioning that 600 images of size 384×256 have been considered. These images are cast into ten classes, each class includes 60 images. Twenty images from each class have been selected as test query images.

The classification rate for each class is the percentage of correct matches of the query images.

The results are shown in Table 3.1. As can be seen, the proposed technique significantly outperforms the conventional histogram and the modified colour histogram known as the annular approach [101] Surprisingly, the annular histogram-based technique shows the poorest results although the number of features used is higher. As illustrated, the proposed technique generally performs slightly better when the YCbCr colour space is used. This is due to the fact that the proposed features mainly represent textures and edges which are well described by the luminance component Y.

Table 3.1 Image classification results

CLASS	Traditional histogram RGB	Traditional histogram YCbCr	Annular YCbCr	Annular CIE L a b	Proposed method RGB	Proposed method YCbCr
Africa	85%	85%	40%	30%	80%	80%
Beach	35%	25%	15%	15%	55%	55%
Building	25%	35%	35%	20%	50%	75%
Bus	30%	50%	35%	5%	85%	90%
Dinosaur	100%	100%	95%	90%	100%	100%
Elephant	65%	45%	5%	10%	65%	60%
Flower	75%	90%	40%	10%	90%	90%
Hours	75%	90%	40%	25%	85%	90%
Snow	95%	5%	5%	10%	25%	55%
Food	1%	50%	15%	35%	35%	60%
Average	58.1%	57.5%	32%	25%	67%	**75.5%**

In order to improve the algorithm of the proposed DCT Histogram quantization [102], the proposed method is further extended and combined with standard deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (3,10)$$

where σ is calculated over the image sub-blocks of YCbCr and HSV colour space of both query image and database image. Then, the outcome is concatenated with the histogram DCT quantization feature, in order to form a new feature of histogram DCT quantization and Standard deviation. The results are shown in Table 3.2. As can be seen, the proposed technique significantly outperforms the conventional histogram and the previous proposed work [102]. Surprisingly, the combined proposed technique performs slightly better when the YCbCr and HSV colour spaces are used based on the classification per class. This is due to the fact that the proposed features mainly represent textures and edges which are

well described by the luminance component Y and the distribution of the colour components.

Table 3.2 Classification results using combined features

CLASS	Traditional histogram RGB	Traditional histogram YCbCr	Proposed method YCbCr [102]	Proposed method RGB [102]	Proposed method YCbCr	Proposed method HSV
Africa	85%	85%	80%	80%	**90%**	**85%**
Beach	35%	25%	55%	55%	55%	50%
Building	25%	35%	75%	50%	55%	65%
Bus	30%	50%	90%	85%	**90%**	**95%**
Dinosaur	100%	100%	100%	100%	**100%**	**100%**
Elephant	65%	45%	60%	65%	60%	55%
Flower	75%	90%	90%	90%	**95%**	**100%**
Hours	75%	90%	90%	85%	**80%**	**95%**
Snow	95%	5%	55%	25%	50%	45%
Food	1%	50%	60%	35%	**60%**	**65%**
Average	58.1%	57.5%	**75.5%**	67%	73.3%	**75.5%**

3.2 **Related works on Content base image retrieval using neural networks**

3.2.1 **Feature extraction**

Extraction of visual content from images, always arise the question is what features to extract that will help perform meaningful retrieval. Different attempts of features were implemented to represent the visual content of the image; these features can be summarized as follows:-

3.2.1.1 **Image sub-blocks**

The image is divided into five sub-images, four equal size regions and one in the middle of the image as shown in Figure 3.1. For each region the DCT transform is applied to obtain DCT coefficients and the first and the second of the colour moment are computed over the

coefficients of the sub-images. All of the computed features are concatenated to form a feature vector for the whole image database.

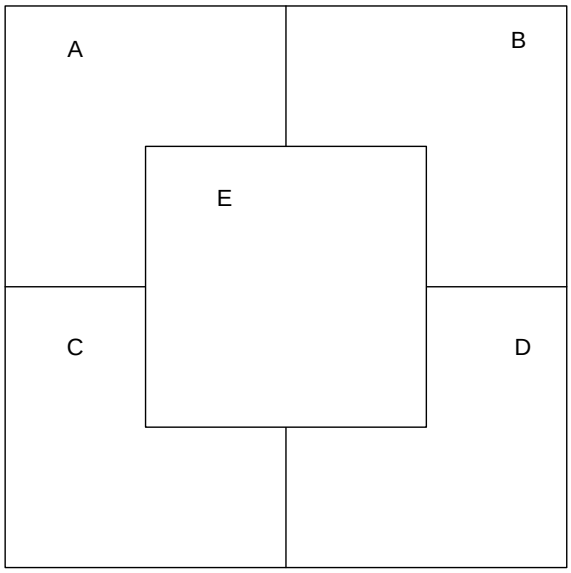


Figure 3-1 Subdivision of an image.

Algorithm ImgSublock
for k=1 to number of images
img_in=convert images from RGB to YCbCr
img_in=divide the img into five region
Compute the 2D DCT F(u,v) of the image sub blocks
Imgdct=DCT2(img_in)
`feature_vector=compute the first and second colour moment`
Concatenate mean and standard division features into array of 1D vectorr

3.2.1.2 **Spatial Approximations Coefficient wavelet sub-band**

The wavelet feature process consists of the following. The query and the database images of size 256×384 are converted to YCbCr representation. The images are wavelet Daubechies decomposed. The four resulting sub-bands are used to extract only the coefficient approximation (CA) values of each colour space. The low-low sub-band (LL) is further decomposed and likewise the four sub-bands at the second level are computed. The

process is repeated up to the fifth level and the outcome sub-band of size 8×12 reorganised into a vector to obtain 96 coefficient values for each colour plane. Hence 288 values constitute the feature vector of the query and the image database for further retrieval processing using neural networks.

Algorithm Approximation Wavelet Cofficient

for k=1 to number of images
img_in=convert images from RGB to YCbCr
Img_in=Obtain the Y component

```
Y=Img_in
[CA CH1 CV1 CD1]=dwt2(Y,'db1');
[CA CH2 CV2 CD2]=dwt2(CA,'db1');
[CA CH3 CV3 CD3]=dwt2(CA,'db1');
[CA CH4 CV4 CD4]=dwt2(CA,'db1');
[CA CH5 CV5 CD5]=dwt2(CA,'db1');
[r,c]=size(CA);
FeatureVector=reshape(CA,1,r*c);
```

end

3.2.1.3 **Non Overlapping Blocks**

The image was divided into non overlapping 32×32 sub-blocks; the resulting sub-blocks are used to compute the first and second order statistics along with the DCT coefficients of the non overlapping sub-blocks. The process is carried out for each colour plane to constitute the feature vector of the image database for further retrieval process using neural networks.

Algorithm Non_overlapping
for k=1 to number of images
img_in=convert images from RGB to YCbCr
Img_in=Divide_img_into 32X32 Non_overlapping
Compute the 2D DCT F(u,v) of the image 32X32 divided blocks
Imgdct=DCT2(img_in)
Compute the mean and standard devision
Concatenate mean and standard division features into array of 1D vectorr
end

3.2.1.4 **DCT_ wavelet sub-bands**

A query and image database of size 256×384 are divided into non overlapping 64×64 sub-blocks; the resulting sub-blocks are further subdivided into seven region wavelet sub-bands decomposed as shown in figure 3.2. Then the results are used to compute the first and second order statistics along with the DCT coefficients of the non overlapping sub-bands. The process is carried out for the luminance Y colour plane to constitute the feature vector of the image database for further retrieval processing using neural networks.

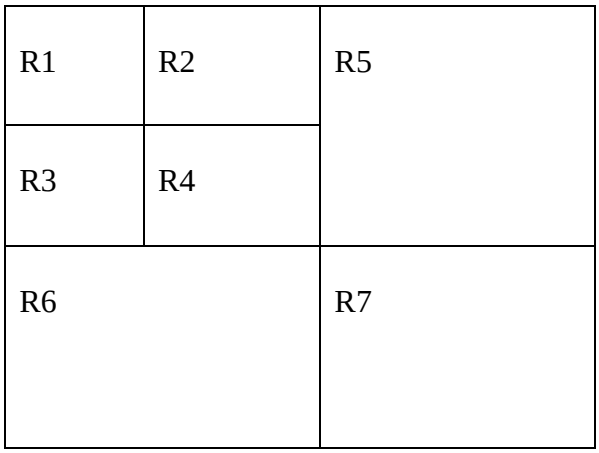


Figure 3-2 DCT block decomposed by factors of two and four.

Algorithm DCT waveletSubband
for k=1 to number of images
img_in=convert images from RGB to YCbCr
Img_in=Divide_img_into seven wavelet Region
Compute the 2D DCT F(u,v) of the image regions
*Imgdct=DCT2(**img_in**$_{(region)}$)*
`feature_vector=compute the mean and SD`
Concatenate mean and standard division of the seven regions features into array of 1D vectorr
end

3.2.2 **Experimental Results**

Extraction of the proposed features was implemented on the Wang database; all the kinds of extracted features are cast into training datasets and testing datasets for each class. A neural network was applied in this work to retrieve the relevant images to the specific class image. This network is a Feed-Forward network with more than one hidden layer. Multiple layers of neurons with nonlinear transfer functions allow the network to learn more complex nonlinear relationships between input and output vectors. As mentioned, the database image features were divided into two sets: a training data set and testing data set, to evaluate the performances of the learning system and in this work, 80% of the available features were selected from each class and used for training while the remaining 20% were used for testing. In practice, the training data set is trained using the neural networks, the training database features selected are fed to the neural networks as input, and the output of the neural networks is an average is applied along the column of the input training data set feature space of the database, where the average outcome is referred or labelled as the output of the neural networks. Once the neural networks are trained the retrieval process is straightforward based on the testing dataset which returns the relevant images per class. The retrieval results are a list of images ranked by their similarity distances to the testing data per class. The similarity distance measure between the feature vectors of testing database image and the training database image is defined in equation 9. Table 4.3 shows the results of the different retrieval performance using neural networks.

Table 3.3 Results of different proposed feature extraction

Class	DCT Wavelet Sub-band	Wavelet Apd6rox on Y	Five sub-images	Non overlapping Blocks
Africa	15%	2%	10%	7%
Beach	24%	16%	12%	17%
Building	20%	14%	13%	10%
Bus	14%	9%	10%	5%
Dinosaur	95%	91%	84%	80%
Elephant	37%	25%	22%	21%
Flower	46%	51%	31%	33%
Hours	64%	37%	20%	41%
Snow	17%	9%	12%	10%
Food	7%	3%	10%	2%
Average	33.9%	25.7%	22.4%	22.6%

As can been seen the result from the table 3.3 is reasonable, but it not satisfactory due to selection of features and the complexity of background of some classes, in addition the colour resolution similarity between some classes for example, food and some faces painted in the African class.

In the following algorithm we address the problem of the poor result, DCT quantization of the coefficients is proposed in the image classification as representative feature which is described in section 3.2.1. Furthermore we also used in the following section in the term of content based image retrieval to resolve the problem of selecting the suitable feature that represent the content of the image. Significant result have been achieved which can be seen in the table 3.4 in the column DCT only.

3.3 Efficient content-based image retrieval scheme based on Semantic Object Detection (SOD)

Image retrieval systems can be cast into two categories: text based and content based systems[87, 94]. Text-based retrieval is a method that manually annotates images by text. However, this method has many limitations because it is highly labour-intensive, time consuming, and impractical with large databases. In content-based image retrieval (CBIR), the main idea is to extract low level features, such as colour, texture, and shape, which best describe the image visual content properties[88, 94, 103-105]. These features are then used to measure similarity. CBIR offers an attractive way to overcome the problems of text-based image retrieval.

In the literature, there has been a growing body of research on CBIR[87, 103]. In [106], the authors represented texture by the first and second order statistics of transform coefficients obtained via Gabor filters. Swain and Ballard have demonstrated the potential of using colour histograms for colour image indexing [104]. In [107] new texture features extracted from spatial blocks of the pre-processed image, BDIP (block difference of inverse probabilities) and BVLC (block variation of local correlation coefficients), have been adopted for CBIR. The authors have also adopted these features in the wavelet transform domain in [108]. In order to improve the retrieval accuracy of content-based image retrieval systems, research focus has been shifted from designing sophisticated low-level feature extraction algorithms to reducing the ‘semantic gap’ between the visual features and the richness of human semantics [88]. In [109], object ontology has been used to define high-level concepts for scenery image retrieval.

In the work described here, an efficient content-based image retrieval scheme based on Semantic Object Detection (SOD) is proposed. The feature extraction process uses quantized discrete cosine transform (DCT) blocks where the retrieval of the query image is performed using a histogram-based similarity measure. SOD aims to reduce the size of the database from which the retrieval of similar images is conducted.

3.3.1 **Semantic Object detection**

The difficulty in content-based image retrieval is how to summarize the low-level features into high-level or semantic descriptors to facilitate the retrieval procedure. One of the most important challenges and perhaps the most difficult problem in semantic understanding of media is visual concept detection in the presence of complex backgrounds. Many researchers have looked at classifying whole images, but the granularity is often too coarse to be useful in real world applications. It's typically necessary to find the human in the picture, not simply global features. Another limiting case is where researchers have examined the problem of detecting visual concepts in laboratory conditions where the background is simple and therefore can be easily segmented. Thus, the challenge is to detect all of the semantic content within an image such as faces, trees, animals, etc.

An early simple detector for city pictures was demonstrated by Vailaya et al [110] they used a nearest neighbor classifier in conjunction with edge histograms. Schneiderman and Kanade [111] proposed a face detection system based on component using the statistics of parts. Chua, et al. [112] applied the gradient energy directly from the video representation to detect faces based on the high contrast areas such as the eyes, nose, and mouth. They also compared a rules based classifier with a neural network and found that the neural

network gave superior accuracy. A comprehensive survey on the area of face detection was reported by Yang, et al[27].

Detecting a wider set of concepts other than human faces turned out to be fairly difficult. In the term of searching image over the internet, Lew [113] showed a system for detecting sky, trees, mountains, grass, and faces in images with complex backgrounds. Fan, et al. [114] used multi-level annotation of natural scenes using dominant image components and semantic concepts. Li and Wang [115] used a statistical modeling approach toward converting images to keywords. Rautianinen, et al. [116] used temporal gradients and audio analysis in video to detect semantic concepts. In certain contexts, there may be several media type available which allows for multimodal analysis. Amir, et al. [117] proposed a framework for a multi-modal system for video event detection which combined speech recognition and annotated video.

To this end, the idea that images could be described by high level knowledge-based concepts which in turn can be represented by specific objects. A typical a collection of image objects were classified into several classes per image object. in order to describe a high level concept, and mimic the semantic meaning of object, semantic object detection were applied as described in the following proposed section,

3.3.2 **DCT Proposed based CBIR scheme**

Figure 4.3 shows a block diagram of the proposed CBIR scheme. As can be seen, the system consists of two different parts: DCT-based image retrieval and Semantic object detection (SOD).

DCT-based retrieval is a low level retrieval technique which extracts texture and colour features in the DCT domain while SOD performs indexing of all images in the database in

order to select images which are likely to be similar to the query one. Therefore, DCT-based retrieval and SOD are jointly operating in the proposed scheme.

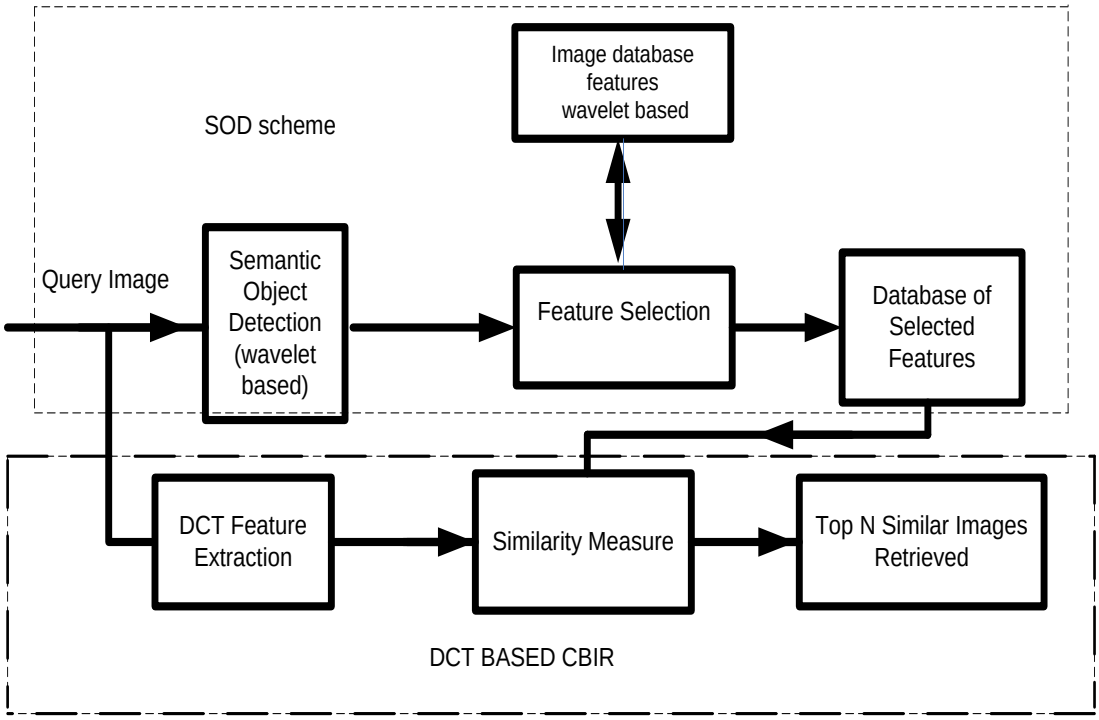


Figure 3-3 Proposed SOD _DCT based CBIR scheme.

3.3.2.1 **DCT-based CBIR features**

The proposed DCT-based feature extraction is described as follows. The input image is first divided into 8×8 non overlapping blocks. Then, each block is DCT transformed. The DC and the first three AC coefficients in Zigzag order are picked as shown in Figure 3.4.

In order to obtain a better quantization of the coefficients that have been selected, normalization is applied by calculating the minimum and maximum values of the entire selected coefficients over the whole database. These are denoted by DC_{min} and DC_{max} the minimum and maximum values of DC coefficients for the whole database and by L the number of quantization bins as described earlier in equations 3.1 to 3.4.

In the experiments *L* is set to 100 based on empirical assessment. Notice that this quantization puts each coefficient in a specified range.

The histogram-based feature vector obtained from quantized DCT coefficients is constructed for both query and database images using *L* levels as described in equations 3.5 to 3.8.

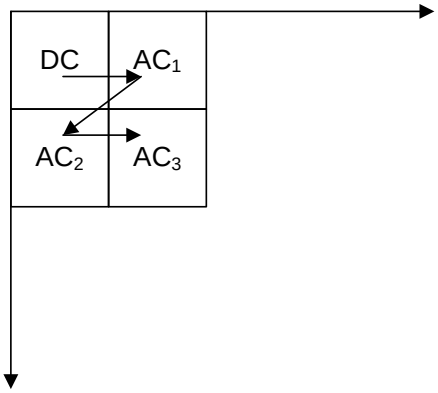


Figure 3-4. Zigzag order used to select DCT coefficients

The similarity measure between a given query image and the images from the database is borrowed from[104]. Let V_{image} and V_{query} be two feature vectors computed as described earlier from a given image from the database and a query one, respectively. The intersection histogram distance is given by

$$H\left(V_{image}, V_{query}\right) = \frac{\sum_{j=1}^{n} \min\left(V_{image\ (j)}, V_{query\ (j)}\right)}{\sum_{j=1}^{n} V_{query\ (j)}} \qquad (3.11)$$

where $V_{image}(j)$ is the j^{th} feature and n represents the number of features. Observe that the larger the similarity distances the better the fit.

3.3.3 **Semantic object detection (SOD) scheme**

SOD exploits the idea that images could be described by high level knowledge-based concepts which in turn can be represented by specific objects. Therefore, a number of object templates are used to represent each concept. The semantic object detection process

consists of the following. Colour images are used in RGB representation. First, the template image is wavelet decomposed. The four resulting sub-bands are used to compute first and second order statistics. The low-low sub-band (LL) is further decomposed and likewise the statistics of the new four sub-bands at the second level are computed. The process is repeated up to the third level to obtain 24 statistical values for each colour plane and hence 72 values constitute the feature vector of the template for semantic object detection. To detect the presence of an object given by its corresponding template *t* in an image, the image is first decomposed into overlapping blocks B_k where the size of each block is equal to that of the template. Then, each block is wavelet decomposed to extract its feature vector V_k which will be compared to that of the template V_t using the Canberra distance, a similarity measure which is given by:

$$d_{t-k} = \sum_{i=1}^{P} \frac{|V_k(i) - V_t(i)|}{|V_k(i)| + |V_t(i)|} \qquad (3.12)$$

where $V_k(i)$ is the i^{th} component of feature vector V_k and p is the number of features. The distance between the template and the image, $d_{t\text{-}image}$, is taken to be the minimum distance of all computed distances as follows:

$$d_{t-image} = min_{k \in \{1,2,\ldots..,p\}}(d_{t-k}) \qquad (3.13)$$

All images in the database are indexed with three different values $X = \{0,1,2\}$. An object is said to be present in a given image if it gives a distance from the templates representing the semantic concept characterised by this object smaller than a given threshold T_2. The image is then labelled by 2. On the other hand, if the distance is larger than a given threshold T_1, the object is not present in the image and will be labelled by 0. If the distance is between

T_1 and T_2, the image is labelled by 1 which means that the object might be present in the image. This process is illustrated in Figure 3.5

$$x = \begin{cases} 0 & non\ object \\ 1 & perhaps \\ 2 & object \end{cases}$$

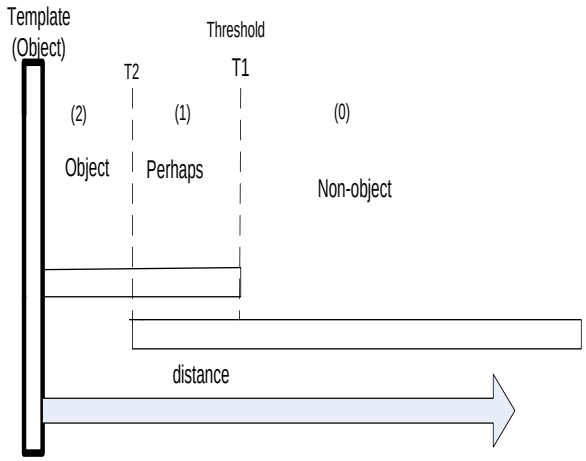


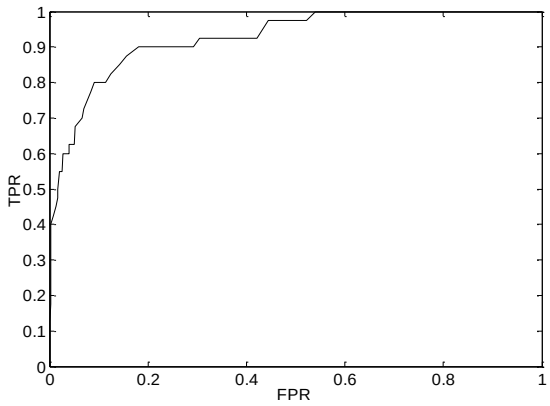
Figure 3-5 Semantic image indexing.

Once the images are indexed, the retrieval process using the proposed CBIR technique described in Section 3.4.1 will be straightforward. Indeed, if the query image is indexed by 0 with respect to a given object, only those images with the same index or with index 1 can be considered by the CBIR technique and likewise, if the query image is indexed with 2. All images with index 0 will be excluded by the CBIR scheme.

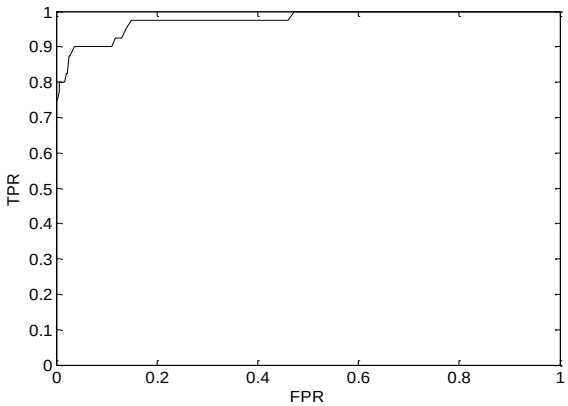
3.3.4 **Experiment results**

The Wang database [100] has been used to evaluate the performance of the proposed CBIR technique. The data set consist of 800 colour images equally divided into 10 different concepts (classes). From each class, 40 images were used as queries and the remaining 40 images for retrieval. In the first set of experiments, the performance of SOD was assessing on the current database. The following five objects have been considered {face, bus, elephant, horse, and building}. ROC curves plot the probability of correct detection (true

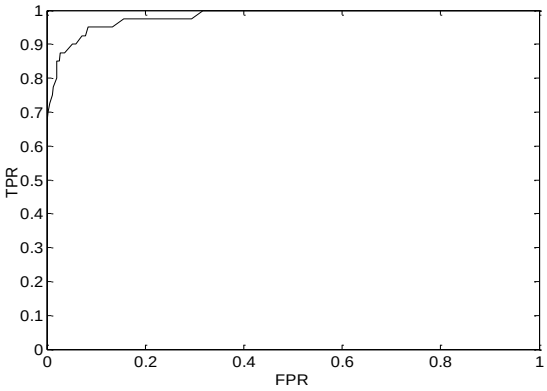
positive rate TPR) against the probability of false detection (false positive rate). Figure 3.6 shows ROC curves for different objects. As can be seen, the performance demonstrates the efficiency of the proposed SOD technique. This also shows that SOD can be used to enhance CBIR.



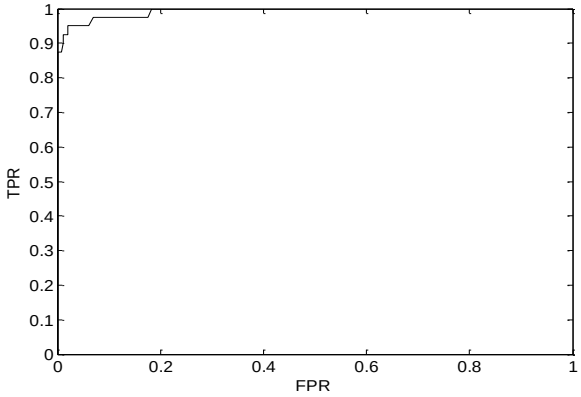
(a)



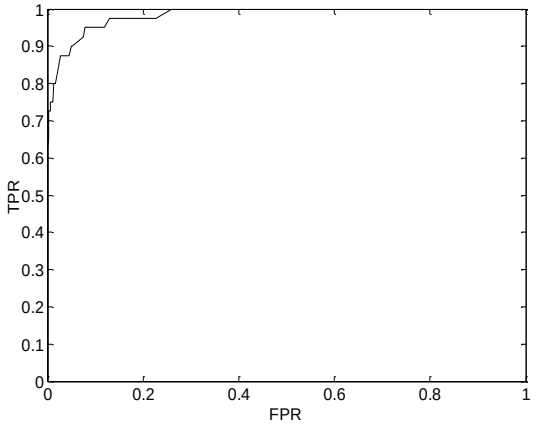
(b)



(c)



(d)



(e)

Figure 3-6 Semantic object detection results (a) African face (b) building (c) Bus (d) Elephant (e) Horse

Table 3.4 depicts the retrieval rate results for various classes with different retrieval techniques. DCT-only refers to the proposed technique without SOD described in section 3.2.2.1 while DCT with SOD refers to the proposed technique described in 3.3.3. As shown, the overall average of the proposed scheme outperforms DCT-only. The retrieval enhancements attributed to the use of SOD are also noticeable if we take DCT-only as a reference point.

Table 3.4 average retrieval rate with different techniques

	DCT only		**SOD_&_DCT (proposed)**	
CLASS	**Top 15**	**Top 20**	**Top 15**	**Top 20**
African	**55.17%**	**51.13%**	**55.5%**	**51.63%**
Beach	**42%**	**38.75%**	**42.00%**	**39%**
Building	**45.83%**	**42.75%**	**46.5%**	**43.62%**
Bus	**69.17%**	**67%**	**80.83%**	**78.12%**
Dinosaur	**96.50%**	**96.62%**	**97.5%**	**97.75%**
Elephant	**53.50%**	**49.25%**	**53.50%**	**49.38%**
Flower	**78.83%**	**73.75%**	**79.5%**	**75.25%**
Horse	**94.33%**	**92.62%**	**94.33%**	**92.62%**
Snow	**37.5%**	**35.5%**	**38.67%**	**36.%**
Food	**65.17%**	**62.13%**	**67.33%**	**64.13%**
Average	**63.80%**	**60.95%**	**65.87%**	**62.75%**

Following to the unsatisfied results mentioned in section 3.2.2. Table 3.5 shows the results of retrieval obtained using the DCT features only proposed in section 3.3.2.1 provided for the sake of comparison. In most of the categories our proposed method has performed better performance than other systems [118-121]. The results are considerably improved by considering the histogram-based feature vector obtained from quantized DCT coefficients.

Table 3.5 Results of proposed CBIR with existing works

Class	SIMPLisity [118]	Histogram Based [119]	Edge Based [120]	Harris corner Detector [121]	Color point [121]	Proposed DCT only
Africa	48	30	45	40	48	55.17%
Beach	32	30	35	31	34	42%
Building	35	25	35	32	33	45.83%
Bus	36	26	60	44	52	69.17%
Dinosaur	95	90	95	92	95	96.50%
Elephant	38	36	60	28	40	53.50%
Flower	42	40	65	58	60	78.83%
Hours	72	38	70	68	70	94.33%
Snow	35	25	40	32	36	37.5%
Food	38	20	40	44	46	65.17%
Average	47.1%	36%	54.5%	43.7%	51.4	63.80%

3.4 **Summary Conclusion**

This proposed system presented in section 3.2 is an efficient content based image classification technique using histogram quantization extracted from DCT coefficients over the all image blocks, the DC and the first three AC coefficients in zigzag order system over all the DCT coefficient blocks. The Euclidean distance is used to measure the similarity in order to retrieve the closest images in the database by applying the KNN classifier. In addition to the simplicity of the proposed technique in terms of computational cost, the experimental results have shown high classification rate with a significantly better

performance in comparison with the conventional histogram and the modified colour histogram annular approach based classification techniques [101].

In the related works proposed system on content based image retrieval, different feature was extracted in order to perform content based image retrieval using cascaded neural networks. The extracted features listed as follows; image sub-block, spatial approximation coefficient sub-band, non overlapping blocks, and DCT wavelet sub-bands.

The training database features selected is fed to the neural networks as input, and an average is applied along the column of the input training feature space of the database which was referred as the output of the neural networks. The experiment results shows promising performance average rate per class of content based image retrieval due to the following reasons such as (1) the complexity of the image database background.(2) the difficulty of applying segmentation object due to the nature of object image classes (3)the false match due to the colour similarity of objects per classes for example between the food class and the African faces class in the Wang database.

Following to the unsatisfied results mentioned in section 3.3.2. Table 3.5 shows the results of the proposed content based image retrieval obtained using the DCT histogram quantization (DCT features only) provided for the sake of comparison. In most of the categories our proposed method has performed better performance than other literature systems.

The proposed method described in section 3.4 is an efficient content based image retrieval technique based on semantic object detection. The idea relies on the fact that existing CBIR schemes could be improved by using further high level knowledge to filter the database for retrieval of similar images. It has been shown through experimental results that the

proposed scheme of the DCT only is outperforms better than other literature techniques, furthermore, on top of that it can be seen from the table result the that Semantic object detection (SOD) proposed scheme by integrate DCT and SOD is outperforms both DCT-only retrieval technique and another related technique from the literature.

Chapter Four

4 Semantic based image retrieval

Neural networks are often used to learn high level concepts from low level features. In the training process need dataset of features either low level features or segmented regions that characterize the image content categories are fed into the neural networks classifier to establish classification in order to link low features of an image with the semantic category label that represent the high level feature.

To solve the semantic gap problem an efficient way is to develop a classifier to identify the presence of semantic image components that can be derived from low level features connected to semantic descriptors. Human face is one among various semantic objects, and it's very important semantic content [122], which is usually also the most concerned centric element in many images and photos. The presence of faces can usually be correlated to specific scenes with semantic inference according to a given ontology. Therefore, face detection can be an efficient tool to annotate images for semantic descriptors.

4.1 Face Detection in DCT Domain

Due to recent advances in broadband networks, image/video compression standards (MPEG) and consumer electronics, image/video data now ranges from simple home albums, videos and movies to online images and videos shared via the World Wide Web. The huge amount of image data generated daily makes it imperative to index the data in a way that enables fast content-based search and retrieval. This has resulted in active research into developing efficient content based image indexing and retrieval technologies. Typically, such indexing or retrieval techniques are based on features such as histograms,

colour, texture, etc. However, these low-level features do not allow for content-based semantic search and retrieval of image data of interest.

Human faces often constitute the most important content in image or video sequences and in such cases the detection and classification of faces is crucial.

Face detection has received increased attention in recent years, and is the first step in many applications such as face recognition, facial expression analysis, content based image retrieval, surveillance systems and intelligent human computer interaction. The performance of these systems depends on the efficiency of face detection. A comprehensive survey of face detection has been given in references [18, 27]. A number of face detection algorithms have been proposed in the pixel domain [27]. To this end Roughly these algorithms can be classified into two groups as follows:-

1. A face pattern is considered as a set of facial features such as eyes, mouth and nose with positions and sizes within an oval shaped area. The presence of a face is concluded from the integration of several detection results [123, 124]. The advantage of these component-based approaches is that the patterns of the components (eyes, nose, oval shape, etc.) might vary less under pose changes, orientation and viewpoint changes than the patterns belonging to the face as a whole. However, it is hard to choose a set of facial features and model their geometrical configuration in such component-based approaches.

2. A face can also be considered as a single pattern and features are extracted from the entire face region. Many methods have been applied such as Gaussian Mixture Distribution Model [36], neural networks [34, 36], principal component analysis (PCA) [125] and support vector machine (SVM) [126].

Little work has been done to address the problem of face detection in the DCT domain. Wang and Chang [40] proposed face region detection directly without decoding of the compressed video sequence, they combined chrominance, shape and DCT frequency information in order to achieve high speed face-detection. Faces are then detected based on colour information as a major detection clue. In their method AC coefficients were used to reduce the number of false detections In addition, they also use some texture information by grouping the DCT parameters into bins and evaluating the energy distribution patterns based on bin statistics. Luo and Eleftheriadis [20], proposed face detection using DCT coefficients based on both colour and texture information. The processed colour information is similar to that done in [40]. Statistical model training and detection formed the bases for texture analysis. In [127], Zhao et al. proposed a DCT-based system that allows the extraction, tracking and grouping of face sequences in MPEG video.

4.1.1 Face detection based on skin colour in image by neural networks

Skin colour information is a very popular and effective feature used for face detection. Although different people have different skin colours, several studies have shown that human faces have a special colour distribution and basic differences are based on their intensity rather than their chrominance [128, 129]. Among feature based face detection methods, using skin colour as detection cue have gained strong popularity. Colour allows fast processing and highly robust to geometric variation of face pattern. The obvious advantage of this method is simplicity of skin colour detection rules that leads to construction of a very rapid classifier.

DCT coefficients as feature based components reduce spatial redundancy and capture compact information about patterns. The skin colour model is created in the YCbCr colour

space, and the reason for choosing the Cb and Cr part of the colour space is it contains chrominance information only and no information about luminance. Thus, skin colour models can be derived from Cb and Cr. By using threshold techniques, skin colour pixels are identified by the presence of a certain set of Cb and Cr values which correspond to the respective ranges R_{Cb} and R_{Cr} values of skin colour. Otherwise, the pixel is classified as non skin colour.

Figure 4.1 shows the proposed system designed with three main categories of operations: pre-processing, feature extraction, and classification using neural networks

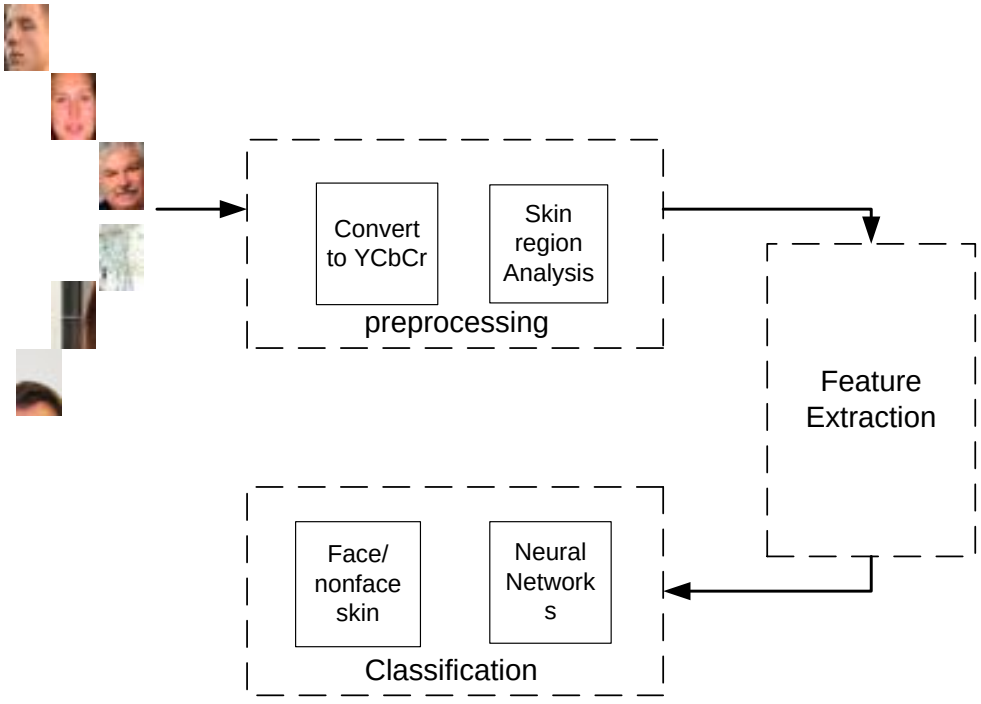


Figure 4-1 the proposed system of Face detection

4.1.1.1 **Pre-processing**

Due to the advantage of skin colour, that is insensitive to face orientation, and occlusion, skin colour is widely used to identify the potential face region. Furthermore, it's used as a pre-processing step to filter out the non facial regions. The processing of skin colour is in fact faster than processing other facial features. Literature survey [27] shows that the YCbCr colour space is one of the successful colour spaces in segmenting skin colour accurately, because mainly the chrominance components are almost independent of luminance component, where Cb and Cr represent the chrominance component. in this work presented the data set of faces and non-faces has been collected and pre-processed by cropping or cutting manually skin face and non-skin face regions. So the Cb and Cr colour space was used to extract DCT coefficient features from their blocks.

4.1.1.2 **Skin colour segmentation**

Skin colour information provides very important features for much research; however the accuracy of skin colour detection is important for face detection [130].

In the proposed method, the image is converted from RGB to YCbCr colour spaces, because RGB is sensitive to the variation of intensity. Many skin detection methods ignore the luminance component of the colour space, in order to achieve models independent of the differences in skin appearance that may arise from the different human races, and also reduce the space dimension. Furthermore, the goal is to model the skin tone, which is more controlled by the chrominance than luminance coordinates. Another argument for ignoring luminance is that skin colour differs from person to person mostly in brightness and less in

the tone itself. The illumination conditions clearly affect the colour of the objects in the scene.

A sample of different human faces were collected and analyzed to obtain histogram distribution of skin colour chrominance component values to represent the likelihood of a pixel belonging to the skin region range. It was found the chrominance component of the skin colour falls in a certain range. The colour is projected on the C_b and C_r plane is inside a predetermined rectangle $C_b \in R_{C_b}$ and $C_r \in R_{C_r}$ i.e., $C_{r1} \leq C_r \leq C_{r2}$ and $C_{b1} \leq C_b \leq C_{b2}$., where $R_{c_b} = [C_{b1}, C_{b2}]$ and $R_{c_r} = [C_{r1}, C_{r2}]$,which is found experimentally to eliminate quickly non-skin face colour, and also improves the segmentation of skin colour regions ranges. The threshold is applied in C_b and C_r colour space as shown in Figure 4.2. Figures 4.3 and 4.4 show the distributions of skin colour sample histograms.

(a) (b)

(c) (d)

(e) (f)

Figure 4-2 Skin face region segmentation (a) RGB image (b) Y component (c)) Cb component (d) Cr component (e) Region segmented on Cb (f) Region segmented on Cr.

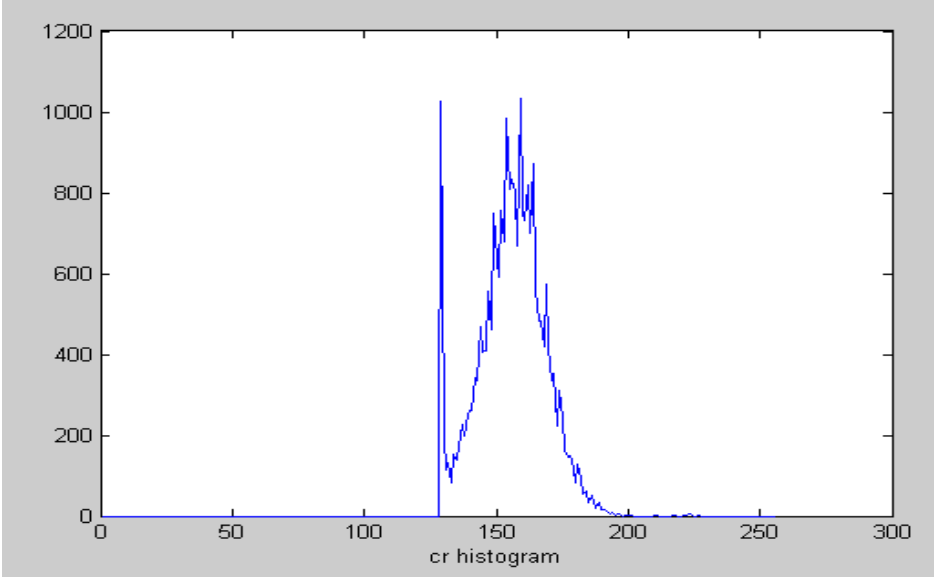


Figure 4-3 Cr histogram distribution sample.

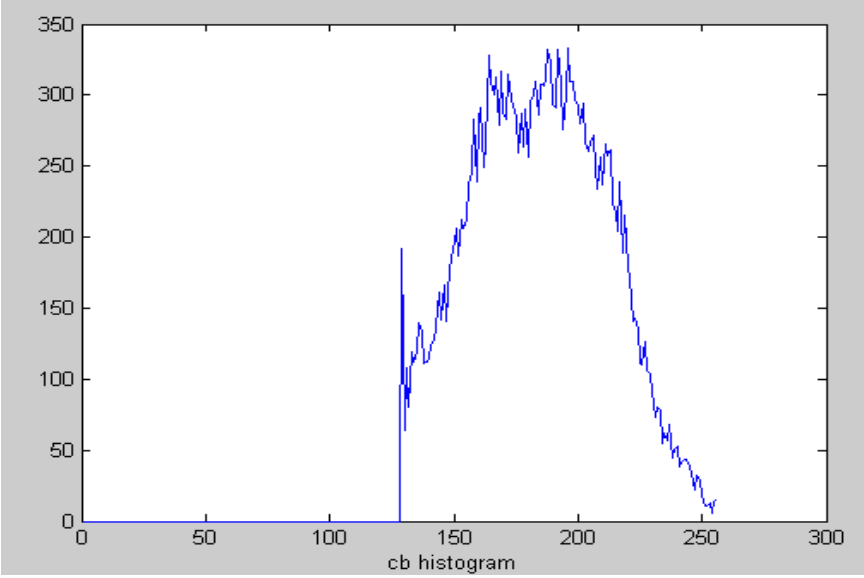


Figure 4-4 Cb histogram distribution sample.

4.1.1.3 **Feature extraction**

The discrete cosine transform (DCT) is used widely in many applications and commonly used in the compressed data domain where it forms the well known baseline JPEG image compression format. Jiang et al. [131] introduced simple, low cost and fast algorithms that extract dominant colour feature directly from the DCT rather than the pixel domain. The extracted DCT coefficients can be used as type of signature which might be useful for recognition tasks, such as facial expression recognition [132].

The proposed technique was derived from [131], where each colour space Cb and Cr is divided into non-overlapping 8×8 pixel blocks. The system calculates the 2D-DCT for each block coming out of the previous stage. Empirically, the upper left corner of the DCT coefficients contains the most important values, because they correspond to the low-frequencies and retain enough information. The uppermost coefficient is called the DC coefficient and corresponds to the average light intensity of the block. The others are called AC coefficients, and provide useful information about the texture detail in the blocks. To form a 1D sequence, a zigzag order applied to each colour space. For each block, the DC coefficients and the first three zig zag order AC coefficients are chosen as a set of 1×4 feature vector coefficients for each colour space sub-block. The output vectors for both Cb and Cr are combined to form a feature vector 1D array that represents the image feature as shown in Figure 4.5.

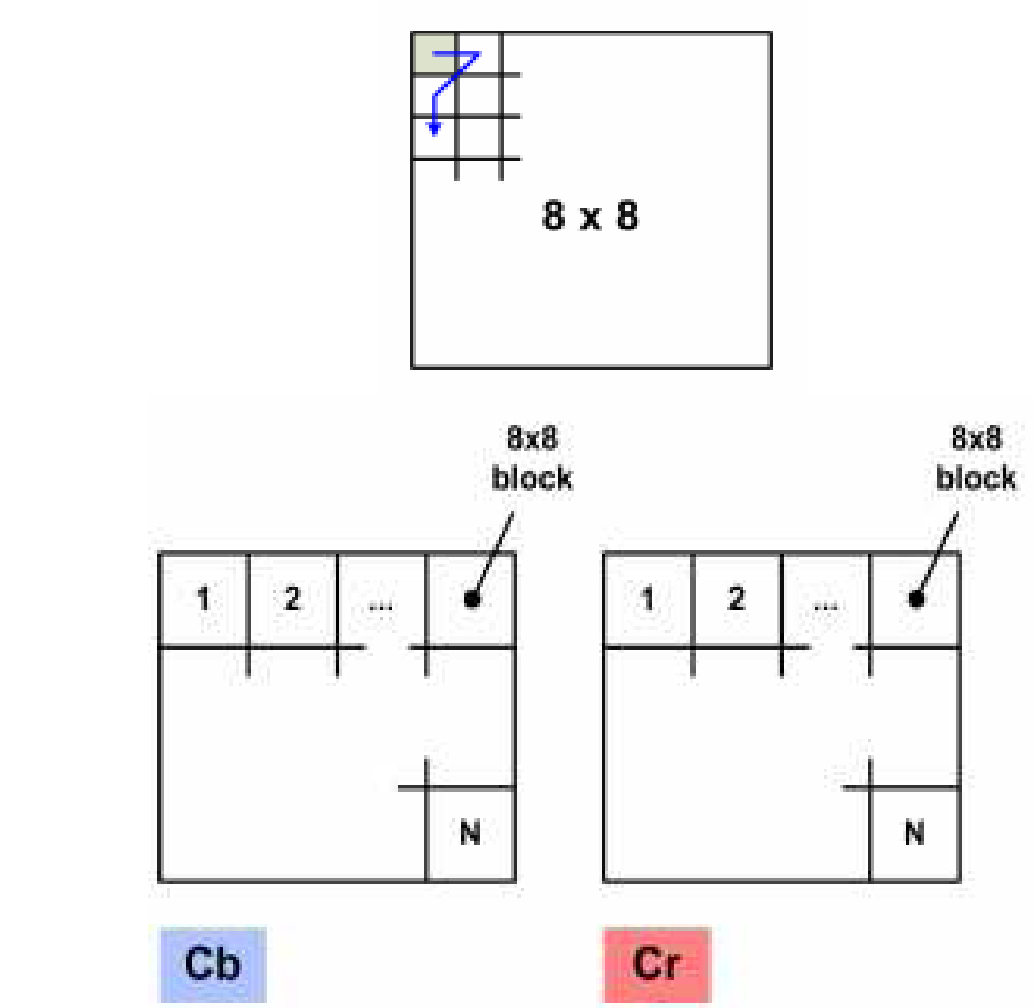


Figure 4-5. Feature extraction from DCT coefficient

4.1.1.4 **Classification**

Neural networks (NN) are often used in face detection. Rowley, Baluja and Kanade [34] proposed a face detection method based on neural networks, that could discriminate between faces and non faces in a large dataset of images. In this work described here, a

multi layer perception (MLP) back propagation neural networks. Features extracted using the Discrete Cosine Transform (DCT) .where the feature images is divided into training dataset and testing data set. NN was applied to learn and training the data set in order to classify them based on the target output labelled by 0.9 for face skin colour and 0.1 refers to non face skin colour. Training is performed using a feature vector obtained from training data set of size 18×27 pixels for true oval faces, which usually means that only face or non face coefficients contribute to the input neural networks. The classification output produces an output of 0.9 for the skin face colour and 0.1 for the non-skin face colour. Input samples and desired targets are repeatedly presented, comparing the output with that desired output and measuring the error and adjusting the weights until sufficient correct output for every input is achieved. The main advantage of choosing Back-propagation neural networks is the simplicity and capability in supervised pattern matching.

4.1.2 **Face detection based neural networks using robust skin colour segmentation**

During previous years, face detection algorithms based on skin colour information have come to the attention of many researchers. Therefore the accuracy of skin colour detection is important to many face detection systems. Kwok et al. [133] proposed an efficient colour compensation scheme for skin colour segmentation. An adaptive skin colour filter was proposed in [134] for detecting skin colour region in colour images. However, it failed to detect skin colour regions when the input image is composed from different human races. Hwei et al. [130] proposed a human face detection based on skin colour segmentation and neural networks. They search for regions where faces might exist, using a skin colour map.

Previous studies have found that pixels belonging to skin regions exhibit similar chrominance components within and across different human races. Mohamed et al. [135] proposed face detection based on skin colour classified by back-propagation neural networks. Skin colour is modelled using threshold techniques and distribution of histogram of sample skin colour face values, skin colour pixels are identified by the presence of a certain set of Cb and Cr values which correspond to the respective ranges R_{Cb} and R_{Cr} values of skin colour. Otherwise, the pixel is classified as non-skin colour. The system design is divided into three main categories: pre-processing, segmentation, and classification using neural networks.

In the proposed algorithm, colour information is used as the main detection cues, and the skin colour model is created in the level of YCbCr. The proposed algorithm works directly on the DCT coefficient parameters, and adopts a Gaussian model applied on samples of face image in order to obtain the likelihood skin colour that belongs to face regions. Then DCT coefficient from the grey image obtained from the previous stage are applied in order to get features from DCT Domain.

In order to narrow the search and speed up the calculations for detecting the skin face regions, the proposed algorithm combines two methods to achieve a fast and accurate face detection system, which relies on feature based methods and image based methods. The feature based method uses a pre-processing of the image based method and guides the search of image based methods using neural networks that examine the face candidate regions instead of performing a huge search in every part of the test image [136].

4.1.2.1 **Pre-processing**

Before training, each of the training dataset image samples are intensity normalized using the equations (4.1) and (4.2), in order to become uniform for application in the next segmentation stage.

$$\bar{I} = \frac{1}{N}\sum_{j=1}^{N} I_i \quad (4.1)$$

$$\hat{I}_i = \left(I_i - \bar{I}\right) + 128 \quad (4.2)$$

Here N is the number of pixels in the training sample, $\bar{I}$ is the average gray value, I_i is the gray value, and $\hat{I}_i$ is the normalized of the gray value of i^{th} pixels.

4.1.2.2 ***Segmentation***

Skin colour information and the accuracy of skin colour segmentation are very important for much face detection research [130]. Many skin detection methods ignore the luminance component of the colour space, in order to achieve models independent of the differences of the skin appearance that may arise from the differences of human races. In this method the image colour space is converted from the RGB to the YCbCr colour space. Moreover, skin pre-processing was used to avoid the exhaustive search for faces, and also to reduce the space dimension. A skin colour was modelled by collecting colour samples from different face images. A total of 97200 skin face pixel samples from 100 colour face images were used to determine the colour distribution of human skin in chrominance blue Cb and chrominance red Cr colour. The skin face samples are filtered using a low-pass filter in order to reduce or remove the noise. The process includes the following steps: in order to

apply the colour distribution for skin colour of different people, the means and the covariance matrix of colour space Cb and Cr is obtained as in equation (4.4).

Colour distribution of skin sample colour space components is clustered in a small area, mainly because the chrominance components are almost independent of luminance component in the space. Although skin color varies from person to person, it tends to get clustered into a compact region in CbCr space.

A maximum likelihood detection scheme is applied in which the image location with highest likelihood is chosen as a face region, fitting the skin likelihood of grey-scale image which represents the likelihood of pixels belongs to skin face. The likelihood for skin face region denoted as skin colour space *x* given its class *k*, where *k* {skin, non-skin}, has a chromatic pair value of (Cb, Cr). The likelihood of skin for this pixel can be calculated by:

$$p(Cb, Cr) = exp[-0.5(x-\mu)^t C^{-1}(x-\mu)] \qquad (4.3)$$

where

$$x = (cb, cr)^t \qquad (4.4)$$

and μ is the mean of both *Cb* and *Cr* and *C* represents the covariance matrix of the two dimensional Gaussian distribution and *N* is the total number of the pixels samples in the face regions. Figure 4.6 shows fitting skin colour using a Gaussian distribution, therefore the colour can be represented using a Gaussian distribution model G = (M, C) where

$$M = (\overline{cb}, \overline{cr}) \qquad (4.5)$$

$$\overline{cb} = \frac{1}{N}\sum_{i=1}^{N} cb_i \qquad (4.6)$$

$$\overline{cr} = \frac{1}{N}\sum_{i=1}^{N} cr_i \qquad (4,7)$$

$$C = \begin{pmatrix} \sigma_{CbCb} & \sigma_{CbCr} \\ \sigma_{CrCb} & \sigma_{CrCr} \end{pmatrix} \qquad (4.8)$$

$\overline{Cb}$ and $\overline{Cb}$ are the Gaussian mean of colour spaces distributions Cb and Cr respectively.

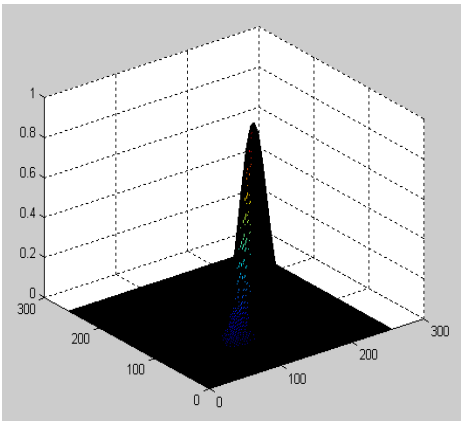
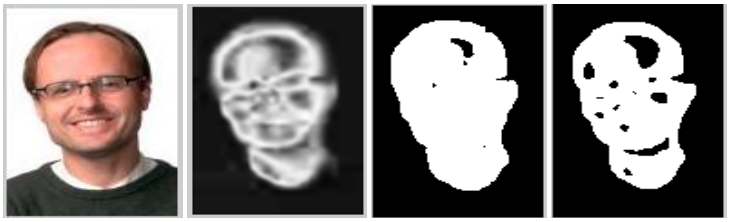


Figure 4-6 Skin colour fitting into Gaussian distribution



(a) (b) (c) (d)

Figure 4-7 (a) Original face image (b) Likelihood skin region (C) Gray skin region (d) binary skin region

In this case the data set image is a frontal face image, moreover, the detected region mainly corresponds to the skin face region and the skin regions are brighter than most other parts of the image. In order to segment the skin region from the rest of the image an adaptive threshold process is applied. The adaptive threshold is based on the observation that stepping the threshold value down may intuitively increase the segmented region. However, the increase in segmented region will gradually decrease (as percentage of skin regions

detected approaches 100%), but will increase sharply when the threshold value is considerably smaller so that other non-skin regions get included. The threshold value at which the minimum increase in region size is observed while stepping down the threshold value will be the optimal threshold. In the program, empirically, the threshold value is decremented from 0.55 to 0.05 in steps of 0.1.

The likelihood skin region is shown as in Figure 4.7 (b) with the original test image in (a), candidate greyscale skin region image in (c) and binary skin regions in (d).

4.1.2.3 **Feature Extraction**

The proposed technique calculates the 2D-DCT for each cropped skin grey-scale block coming out of the previous stage. Empirically, the upper left corner of the 2D-DCT matrix contains the most important values; because these correspond to low-frequency coefficients which contain most energy. The upper left corner coefficient is the DC coefficient corresponds to the average intensity of the block. The others are called AC coefficients which provide useful information about the texture detail in the block. For each block of Cb and Cr, the DCT coefficient features are arranged in zigzag order. Selecting the DC and the first three AC coefficients forms a set of $1 \times N$ vector coefficients feature from both chromatic Cb and Cr. A matrix of $1 \times N$ coefficients obtained for both Cb and Cr colour space components within the processed image block, are merged into one feature vector to be homogenize in order for neural networks to learn and converge as shown in Figure 4.5.

4.1.3 **Experiment and results**

The first algorithm of face detection, two classes of 100 images 50 image per class of face and none face, the images was collected and then normalized to specific size as described in the algorithms. The experiment is carried out applied to unknown input test images

containing a face or non-face. A sliding overlapping window of size 18×27, scans the image, with different overlap parameters 1, 2,.. up to half the width (in this experiments 9 pixel is half of the window). Each part of the unknown test image is scanned and the extracted DCT features are fed as a vector to the neural networks for testing with the training data set have been trained. In order, the unknown test images are classified to see if its contain a face or non face.

The experimental results show that the proposed system is significant and the neural network is able to detect and classify pattern features accurately under different overlap sliding scan window conditions. The convergence response with the training dataset shows accurate and excellent face and non face classification.

In the second algorithm a experimental results are shown to present the performance of the proposed method. The experiment was applied on an unknown input test image containing a face or non-face. Sliding overlapping window scan was applied as mentioned; the features extracted from the DCT are fed to the trained neural networks of the dataset of images in order to classify the input test image as face or non face. The experiment results from the classifier system shows promising results, in that the neural network is able to detect and classify pattern features accurately under different overlapping sliding scan windows over the unknown input test image.

4.1.4 **System Evaluation**

The Jack-knife technique was employed to evaluate the performances of the learning system used in this work with the use of 80% randomly selected samples for training and the remaining 20% for testing. The performance criteria used in the work are accuracy, sensitivity, and specificity which is measured using the common biometric measures,

namely the true positive ratio (TPR) and the false positive ratio (FPR) as explained in chapter 2 section 2.2.1. The most common way to evaluate the performance is the use of Receiver Operating Characteristic (ROC) curves. An ROC curve plots the *FPR* on the x-axis and the corresponding *TPR* on the y-axis such that the positive sloping diagonal line corresponds to random guessing and the system with best performance is the one on the ROC curves which is furthest from the negative sloping diagonal line in the upper-left direction.

Intensive experiments were run on 100 images of face and non face cases using a Neural Network with one hidden layer and with numbers of hidden nodes ranging from 1 to 10. For each number of hidden nodes, 8 results were generated, which were averages over 10 iterations carried out using the Jack-knife technique (80% randomly used for training and rest for testing). The results generated were the number of hidden nodes, TP, FP, FN, TN, Accuracy, Specificity and Sensitivity. Hence 100 learning and testing experiments were carried out for each case.

The results are presented in Figure 4.8 in the form of a ROC (Receiver Operating Characteristic) curve for computing the rate of recognition. This figure shows ROC curves for all the face and non-face features that were computed. The experiment was carried out using 24 input features based on the Jack-knife technique for each CCNN configuration and the average TPR (true positive rate) and FPR (false positive rate) were recorded, At the end of this experiment the best results, obtained for a CCNN in the first Algorithms with 24 input features and 1 hidden layer, were as follows: using 4 and 9 hidden nodes gave the best recognition results which was 90% for accuracy, as well as values of 90% for both of sensitivity and specificity.

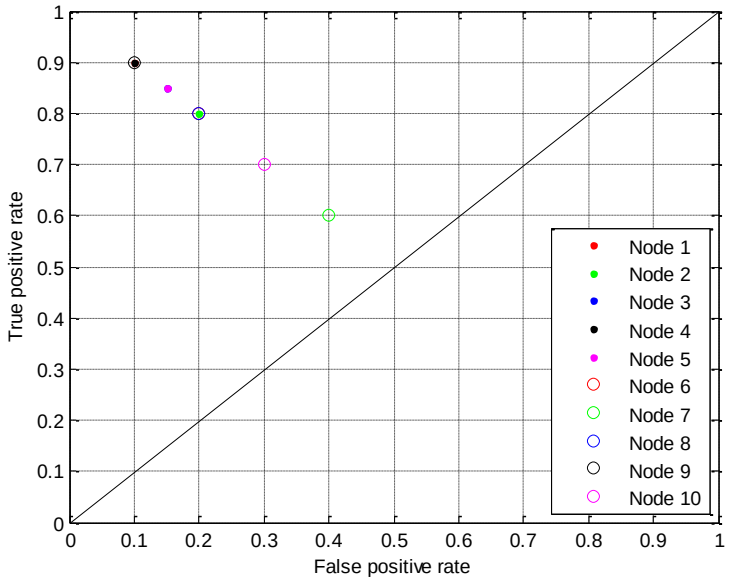


Figure 4-8 ROC graph for average TP and FP values for 24 input feature

The CCNN with 48 input features in the second algorithm, and using 3 and 6 hidden nodes gave 94.5% for accuracy, sensitivity and specificity respectively, as seen in Figure 4.9 which shows the corresponding ROC curve for computing the rate of recognition.

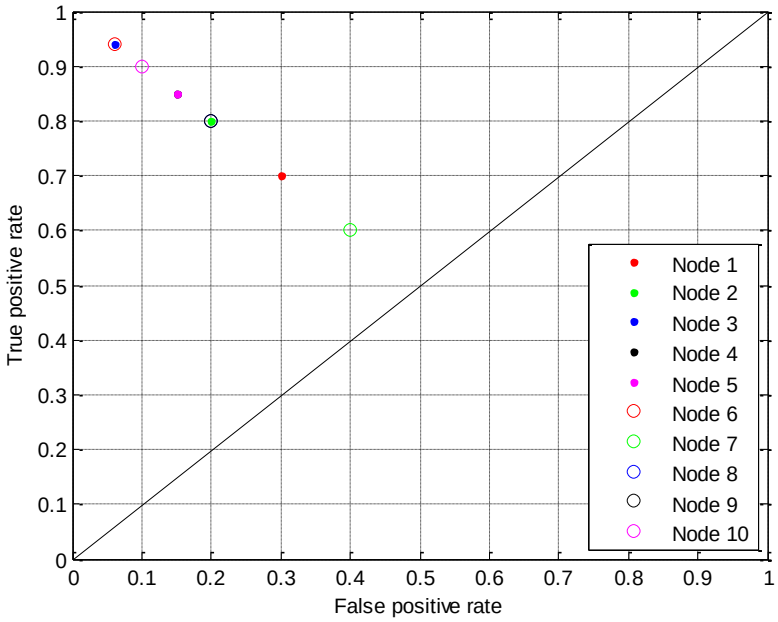


Figure 4-9 ROC graph for average TP and FP values for 48 input feature

Compared with the result achieved, the second scheme outperformance recognition rate for the neural networks classifier as illustrated above.

4.1.5 Content based face image retrieval through DCT coefficients

Content based image retrieval (CBIR) has been a hot research topic over the last decade. A number of image features based on colour, texture, and shape attributes in various domains have been reported in the literature[95, 103]. Recent research is starting to develop image analysis and content feature extraction directly from the compressed domain[96]. CBIR system operation can be classified into two phases: indexing and searching. In the indexing phase, each image of the database is represented by a set of attribute features colour, texture and shape. In the searching phase, when the user selects a query image, a query vector feature is computed. Using a similarity distance measure, such as the well known Euclidian distance, the query vector is compared to the feature vectors in the feature database and retrieves to the user the images that are most similar to the query image.

To provide fast feature extraction from compressed images, a new wave of research effort is to direct access to feature extraction in the compressed domain[9, 12]. All existing research on the compressed domain is limited to the DCT domain. The logic behind this is that the DCT is generally a good approximation to principal component extraction, which helps to process and highlights the signal frequency features, and most existing compressed image data is JPEG and MPEG based. In the proposed method a new simple method of DCT feature extraction is utilized to accelerate the speed and decrease the storage needed in image retrieval aimed at direct content access and extraction from JPEG compressed domain. The method extracts the averages of some DCT block coefficients. This method needs only a vector of the average of the first six coefficients per block sorted by the zigzag

approach over whole image blocks In this retrieval system, for simplicity, images of both query and database are normalized and resized from the original database by taking into account the position of the face based on the centred position of the eyes. The normalized image is uniformly divided into non overlapping 8×8 blocks of pixels. Therefore, each block is associated with a feature vector derived directly from the DCT. Users can select any query as the main theme of the query image. The retrieval of images is based on the relevance between a query image and any database image, the relevance similarity is ranked according to the closest similar measures computed by the Euclidean distance. The experimental results show that with this approach it is easy to identify main objects and reduce the influence of background in the image, and thus improve the performance of image retrieval.

4.1.5.1 **Feature Vector Extraction**

For efficient image feature extraction, the method uses the average of all the DCT coefficients in the compressed domain as the feature vectors. The images, either query or database, are normalized by cropping based on the coordinate of the centred position of the two eyes and then the image obtained is resized. The output image is divided into sub-image blocks (non-overlapping 8×8 blocks): $b(i,j), i = 1, \ldots\ldots\ldots, p \; and \; j = 1, \ldots\ldots\ldots . . q,$ then the DCT is performed independently on the sub-image blocks and the DCT coefficients are coded in a zigzag order. For each sub-block containing one DC coefficient and 63 AC coefficients, are extracted averages over all DCT coefficient blocks, which are the most upper left coefficients, the DC and the first five AC coefficients in zigzag order over all DCT coefficient blocks. The following characteristics are considered: (1) the DC coefficient of each sub-block represents average energy of the image; (2) all the

remaining coefficients within a sub-block contain frequency information which produces a different pattern of image variation; (3) the coefficients of some regions within a sub-block also represent some directional texture information, for example, the coefficients of uppermost region and those of the leftmost region in DCT transform domain represent some vertical and horizontal edge information respectively.

In the work described here, images of both query and database are normalized and resized from the original database based on the centred position of the eyes as shown in Figure 4.10; the normalized image is equally divided into non overlapping 8×8 blocks of pixels. Therefore, each of these is associated with a feature vector derived directly from discrete cosine transform DCT. Users can select any query as the main theme of the query image. The retrieval is the relevance between a query image and any database image as in the block diagram shown in Figure 4.11, the relevance similarity is ranked according to the closest similar measures computed by the Euclidean distance.

Figure 4-10 Original and normalized image to left and right.

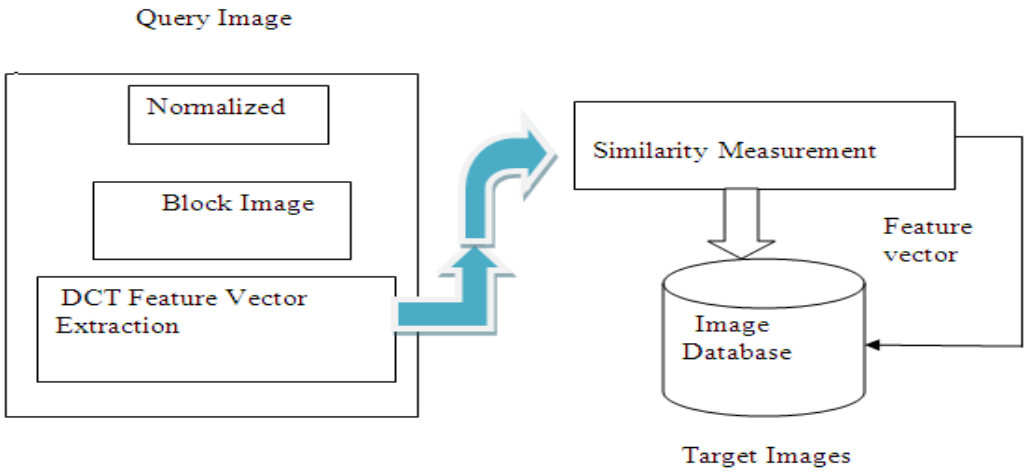


Figure 4-11 Block diagram of proposed retrieval system

4.1.5.2 **Similarity Measure**

Large image database systems mostly require efficient comparisons as well as feature extraction in order to provide reasonable responses to an image query. The similarity measure for a given query image involves searching the database for block DCT vectors similar to the input query. Euclidean Distance is a suitable and effective method which is widely used in the image retrieval area. The retrieval results are a list of images ranked by their similarity distances with the query image. The similarity distance measure between the vectors of query image and the database image can be defined below as described in 4.9.

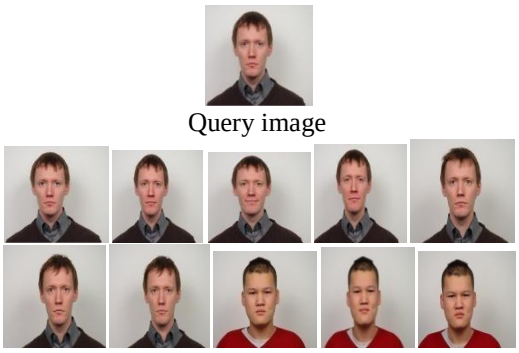
$$D(I_q, I_p) = \frac{\sqrt{\sum_{i=1}^{N}(I_{q_i} - I_{d_i})^2}}{N} \qquad (4.9)$$

where D is the distance between the feature vector I_q and I_d and N represents the number of DCT blocks. The computed distances are ranked according to closeness; in addition, if the distance is less than a certain set threshold, the corresponding original image is close to or a match to the query image.

4.1.5.3 **Experimental Results**

In order to evaluate the retrieval efficiency of the proposed method, it has been implemented on a database of 120 face images, composed of 12 different images per person with different facial situations such expression and positions. It exploits performance measures such as recall and the precision, which are widely used in image retrieval to evaluate the retrieval performance. Recall is the ratio of the number relevant images retrieved to the total number of the relevant images in the database. Precision is the ratio of the number of the relevant images retrieved to the total number of the irrelevant and relevant images retrieved as defined in section 2.4.1.2.1:

The experiment with the proposed method shows promising results based on the above equations. The retrieval results of the proposed method based on the top 10 similar images are shown in Figure 4.12 (a) and (b) shows 58% Recall and 70% Precision.



Query image

Retrieved images

(a)

Query image

(b) Retrieved images

Figure 4-12 Two sets of retrieval results in (a) and (b).

4.2 Conclusion and summary

In the first algorithm proposed in section 4.1.1 a new algorithm for face detection directly from the DCT compressed domain is introduced. DCT coefficient vector features are extracted after assessment of pre-processing of face skin candidate using skin colour information of Cb and Cr colour space, with a back propagation neural networks classifier. The algorithm divides the problem into three stages: pre-processing, feature extraction, and classification using back propagation neural networks. The system has been tested on a dataset of upright frontal colour face images were collected and normalized from the internet and achieved significant detection rates. These methods will improve the detection of faces in compressed images.

The second algorithm proposed in section 4.1.2. A face detection in the DCT compressed domain was carried out, DCT coefficient vector features extracted after segmenting a face skin candidate using a Gaussian mixture model derived from both Cb and Cr colour spaces, along with neural network classifier. The problem is divided into three stages: pre-processing, segmentation, and classification. This schema has been tested on a dataset of upright frontal colour face images normalized from the Internet. Compared with the result achieved, this scheme outperformance recognition rate for the neural networks classifier as illustrated in section 4.1.4.

The algorithm proposed in the section 4.1.5 a simple content based face image retrieval method uses the average over all DCT coefficients blocks of the DC and the first five AC coefficients in zigzag order. The DCT transformation was used to extract the feature vectors directly from the DCT domain of both query image and the database images. Euclidean distance was used to measure the similarity distance in order to retrieve the best closest similar to the query image. Using only the main representative coefficients that are located in the upper left of the DCT blocks not only speeds up the calculation but also reduces the storage space problem. The experimental results show promising results.

Chapter FIVE

5 Hybrid low level and high level based image retrieval

In the literature, there has been a growing body of research on high-level semantics and CBIR [87, 103]. In this regard, texture is a visual feature that refers to innate surface properties of objects and their relationships to the surrounding environment [137]. Texture can provide important information for image classification as it describes the content of many real-world images such as fruit skin, clouds, trees, bricks, and fabric. Hence, texture is an important feature in defining high-level semantics for image retrieval purpose [87]. In the conventional texture features used in CBIR, there are statistical texture features using grey-level co-occurrence matrices (GLCM) [137], edge histogram descriptors (EHD) (which is one of the MPEG-7 texture descriptors) [138], and wavelet moments [139]. Among the various texture features, Gabor features and wavelet features are widely used for image retrieval and have been reported to well match the results of human vision studies [5, 140]. In [107] new texture features extracted from spatial blocks of the pre-processed image; BDIP (block difference of inverse probabilities) and BVLC (block variation of local correlation coefficients) have been adopted for CBIR. The authors have also adapted these features in the wavelet transform domain in [108].

Among the existing research, extensive experiments on CBIR systems have shown that low-level visual features often fail to describe the high-level semantic concepts in user's minds [17] and the main issue is how to narrow the semantic gap between the low-level image features and the high-level semantic concepts.

In order to improve the retrieval accuracy of content-based image retrieval systems, the research focus has been shifting from designing sophisticated low-level feature extraction algorithms to reducing the gap between the visual features and the richness of human semantics [88]. Here, we further extend the algorithm proposed in chapter Three in section 3.3 in order to work on any different CBIR algorithms or any different database image, a new content-based image retrieval scheme based on Semantic Object Detection (SOD) was proposed. The feature extraction process of BDIP and BVLC was borrowed from [107]. SOD aims to reduce the size of the database from which the retrieval of similar images is conducted. The use of SOD has been shown to improve the retrieval performance and outperform related work in the literature.

5.1 CBIR scheme based BDIP_BVLC

Figure 5.1 shows a block diagram of the proposed CBIR scheme. As can be seen, the system consists of two different parts, BDIP_BVLC-based image retrieval and Semantic object detection (SOD).

BDIP_BVLC-based retrieval is a low level retrieval technique which extracts the first and second moments of BDIP and BVLC for each class of the RGB colour component and combines these moments as a feature vector, while SOD performs indexing of all images in the database in order to select images which are likely to be similar to the query one. Therefore, BDIP_ BVLC-based retrieval and SOD are jointly operating in the proposed scheme.

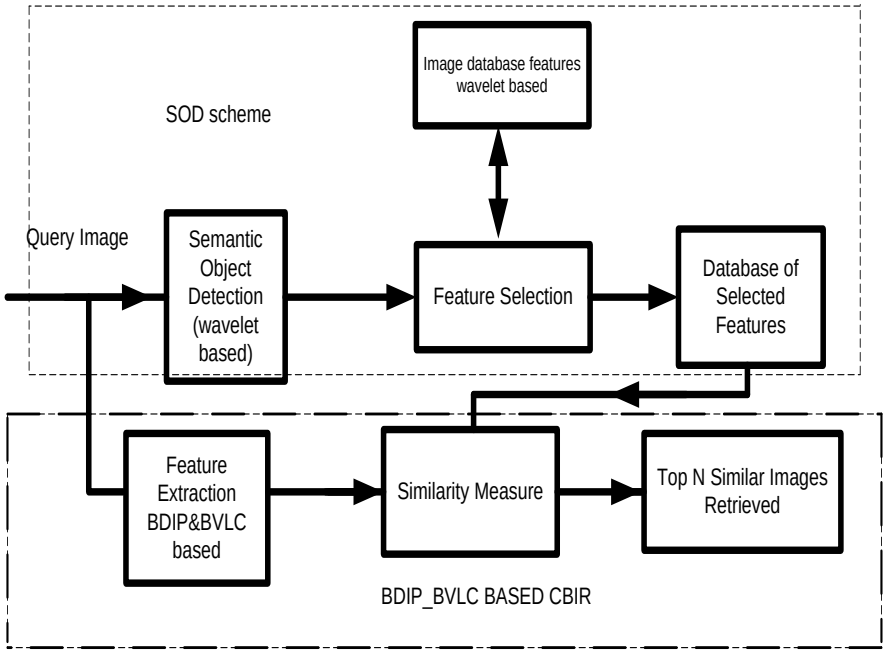


Figure 5-1 The proposed BDIP_BVLC and SOD based CBIR scheme.

5.1.1 **Block difference of inverse probabilities (BDIP)**

BDIP is defined as the difference between the number of pixels in a block and the ratio of the sum of pixel intensities in the block to the maximum in the block. That is

$$BDIP = M^2 - \frac{\sum_{(i,j)\in B} I(i,j)}{Max_{(i,j)\in B} I(i,j)} \qquad (5.1)$$

where $I(i,j)$ denotes the intensity of a pixel (i,j) and B a block of size $M \times M$. The larger the variation of intensities there is in a block, the higher the value of BDIP.

Figure 5.2(a) shows some original images and the corresponding BDIP images whose pixel intensities represent the negatives of BDIP values. The block size was chosen as 2×2 and higher BDIP values are shown darker. In Figure 5.2(b), the insides of objects and the backgrounds are shown bright, while edges and valleys are shown dark. Therefore, the

effectiveness of the BDIP feature for extracting edges and valleys is noticeable. Figure 5.2(c) contains BVLC content of the same images related to the next section.

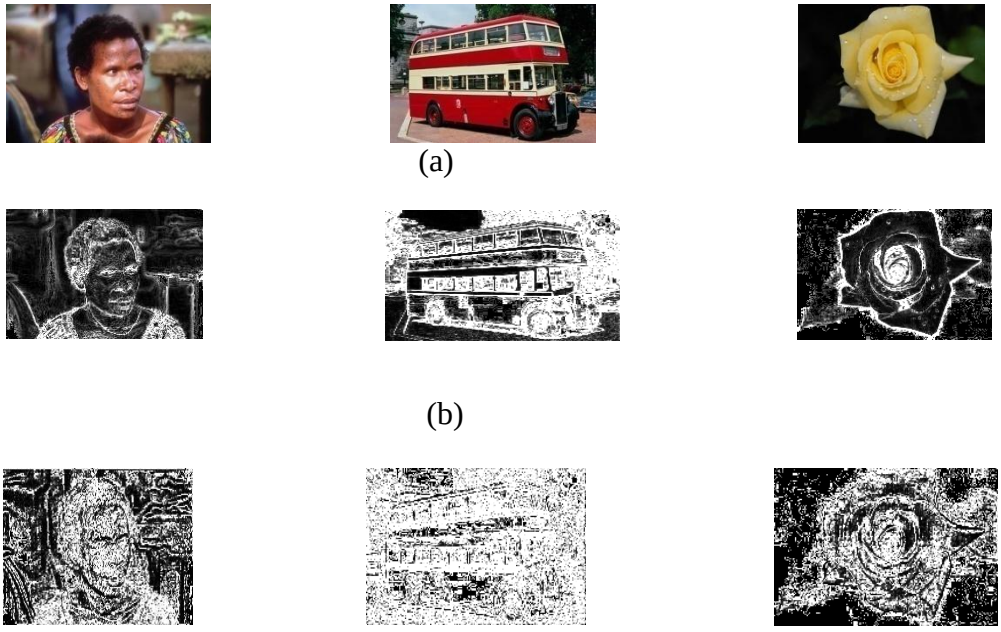


Figure 5-2 (a) Original images (b) BDIP images (c) BVLC images

5.1.2 **Block variation of local correlation coefficients (BVLC)**

The BVLC texture feature is defined as local covariance normalized by local variance. That is

$$p(k,l) = \frac{\frac{1}{M^2}\sum_{(i,j)\in B} I(i,j)I(i+k,j+l) - \mu_{0,0}\mu_{k,l}}{\sigma_{0,0}\sigma_{k,l}} \qquad (5.2)$$

where B is a block of size $M \times M$ and $\mu_{0,0}$ and $\sigma_{0,0}$ denote the local mean and standard deviation respectively, of the block B. The notation (k,l) denotes a pair of horizontal and vertical shifts associated with the four orientations $(-90^0, 0^0, -45^0, 45^0)$. As a result, $\mu_{k,l}$ and $\sigma_{k,l}$ represent the mean and standard deviation of the block shifted by (k,l), respectively.

Figure 5.3 shows pixel configurations for 2×2 windows and the corresponding windows after shifts in each of the four orientations which are required to compute $p(0,1), p(1,0), p(1,1)$ and $p(1,-1)$. As a result, the value of BVLC is expressed as

$$BVLC = max_{(k,l)\in O_4}[p(k,l)] - min_{(k,l)\in O_4}[p(k,l)] \quad (5.3)$$

$$O_4=\{(0,1),(1,0),(1,1),(1,-1)\}$$

Where O_4 is the four orientations $(-90^0, 0^0, -45^0, 45^0)$.

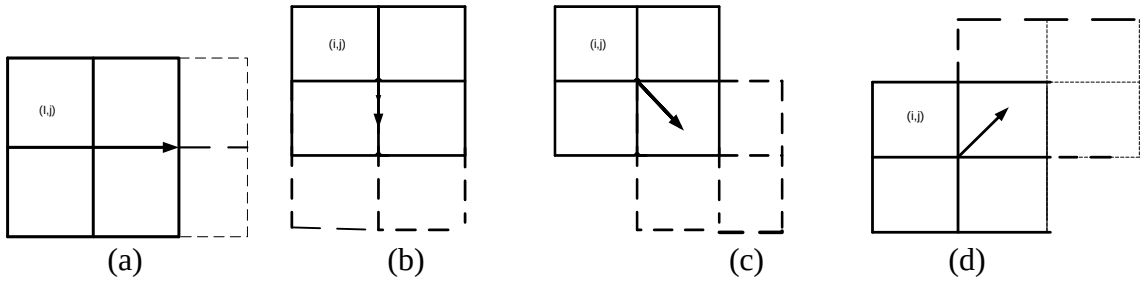


Figure 5-3 Pixel configurations in 2 x 2 windows and their corresponding (a) *P*(0, 1); (b) *P* (1,0); (c) *P*(1, 1); and (d) *P*(1,-1)

In Figure 5.2 (c), it can be seen that the intensity in the BVLC images is determined by texture smoothness.

5.1.3 **BDIP and BVLC based CBIR features**

BDIP and BVLC based feature extraction was borrowed from [107] and described as follows. The image retrieval method was implemented based on the combination of BDIP and BVLC moments. A set of the first and second moments of BDIP and BVLC are used as a feature vector for CBIR.

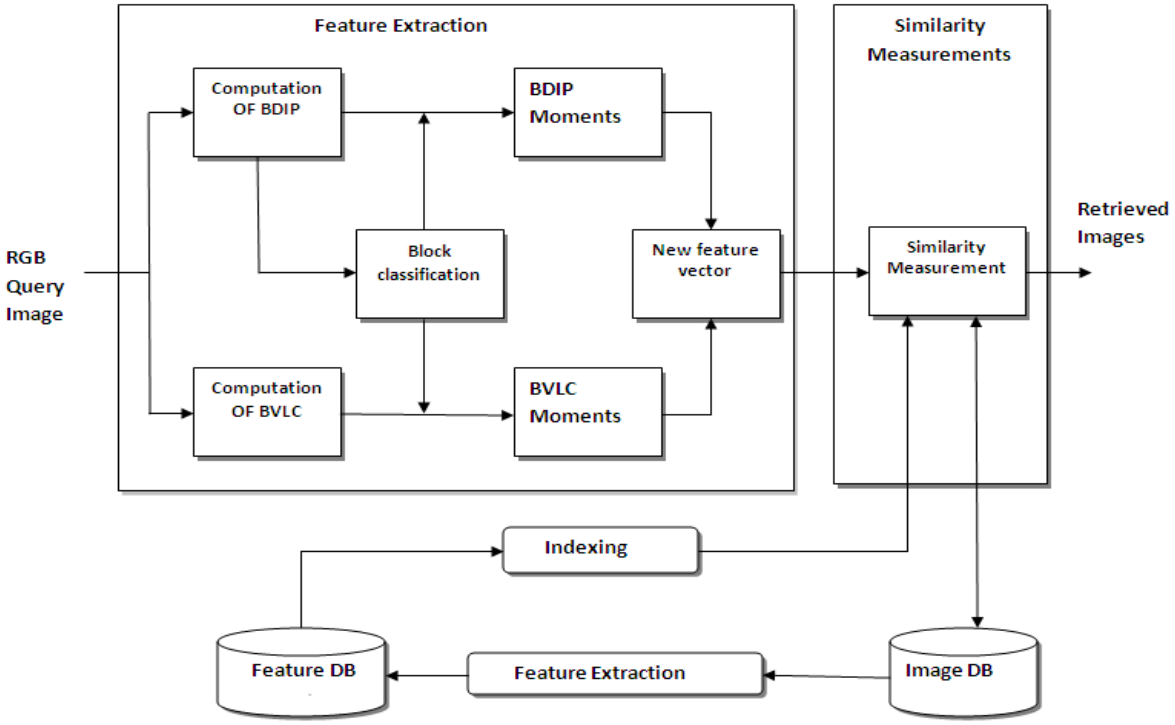


Figure 5-4 . Block diagram of an image retrieval system using the combination of BDIP and BVLC moments

Figure 5.4, shows [107], the block diagram of an image retrieval system using the combination of BDIP and BVLC moments . When a query colour image enters the system, each colour component image is divided into non-overlapping blocks of size $M \times M$; the system then computes BDIP and BVLC in each block and classifies all the blocks into eight classes. The purpose of the classification is to reflect the characteristics of several homogeneous regions or objects, which an image generally contains, using the proposed features [107]. The block classification proceeds as follows. In the first step, all the blocks are classified into two groups, and the average of BDIPs over all blocks in the image is used as a threshold. In the second step, all blocks in each of the two groups are classified again into two groups, but this time the average of BDIPs over all blocks in each group is used as a threshold. In the last step, which repeats the same procedure as in the second step, all the blocks are classified into eight classes. After the block classification, the system

computes the first and second moments of BDIP and BVLC for each class and combines these moments as a feature vector, which can be written as

$$F = [D_{R,}V_R, D_{G,}V_G, D_{B,}V_B] \quad (1)$$

$$D_c = [\mu_c^1(D), \mu_c^2(D), \ldots\ldots\ldots \mu_c^n(D), \sigma_c^1(D), \sigma_c^2(D), \ldots\ldots \sigma_c^n(D)], C \in \{R, G, B\} \quad (5.4)$$

$$V_c = [\mu_c^1(V), \mu_c^2(V), \ldots\ldots\ldots \mu_c^n(V), \sigma_c^1(V), \sigma_c^2(V) \ldots\ldots\ldots \sigma_c^n(V)], C \in \{R, G, B\} \quad (5.5)$$

Where D_c and V_c denote the BDIP and BVLC moment vectors respectively, for each colour component image and $\mu_c^i(.)$ and $\sigma_c^i(.)$ denote the mean and standard deviation respectively for each i^{th} class for each colour component image.

In order to retrieve a given number of the most similar target images the system finally measures the similarity distance between the feature vector of a given query image and the rest of the images in the database using as similarity measure the Mahalanobis distance which is given by:-

$$D(V_q, V_t) = \left(\sum_{i=1}^{n} \left| \frac{V_q(i) - V_t(i)}{\sigma(i)} \right|^m \right)^{1/m} \quad (5.6)$$

Here $V_t(i)$ and $V_q(i)$ are the ith components of the query feature vectors and target feature vectors, m and n denote a metric order and the dimension of a feature vector, $\sigma(i)$ is the standard deviation of the i^{th} component for feature vectors in a feature database.

5.2 Semantic object detection (SOD) scheme

SOD exploits the idea that images could be described by high level knowledge-based concepts which in turn can be represented by specific objects. Therefore, a number of object templates are used to represent each concept. The semantic object detection process is as follows. Colour images are used in RGB representation. First, the template image is wavelet decomposed. The four resulting sub-bands are used to compute first and second statistics. The low-low sub-band (LL) is further decomposed and likewise the statistics of

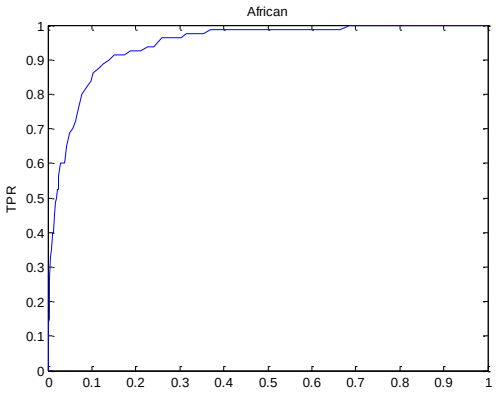
the new four sub-bands at the second level are computed. The process is repeated up to the third level to obtain 24 statistical values for each colour plane and hence 72 values constitute the feature vector of the template for semantic object detection. To detect the presence of an object given by its corresponding template *t* in an image, the image is first decomposed into overlapping blocks B_k where the size of each block is equal to that of the template. Then, each block is wavelet decomposed to extract its feature vector V_k which will be compared to that of the template V_t, using the Canberra distance equation 3.12 and 3.13, as described in section 3.3.3.

All images in the database are indexed with three different values $x = \{0,1,2\}$. An object is said to be present in a given image if it gives a distance from the templates representing the semantic concept characterized by this object smaller than a given threshold T_2. The image is then labelled by 2. On the other hand, if the distance is larger than a given threshold T_1, the object is not present in the image which will be labelled by 0. If the distance is between T_1 and T_2, the image is indexed by 1 which means that the object might be present in the image. This process is illustrated in Figure 4.5.

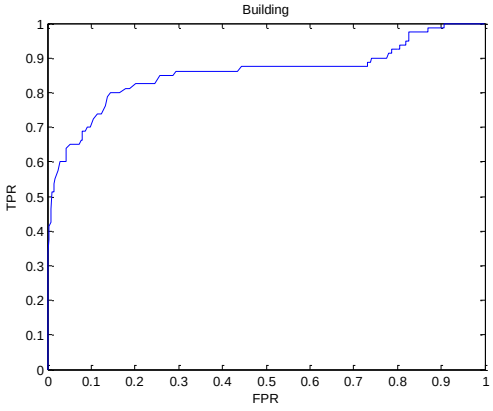
Once the images are indexed, the retrieval process will be straightforward using the adopted method CBIR technique as described in Section 5.1.3. Indeed, feature selection is applied in order to find the new database feature selection. If the query image is indexed by 0 with respect to a given object, only those images with the same index or with index 1 can be considered by the CBIR technique. Likewise, if the query image is indexed with 2, all images with index 0 will be excluded by the CBIR scheme.

5.3 **Experimental results**

In the first experiment, the Wang database [100] has been used to evaluate the performance of the proposed CBIR technique. 800 colour images have been used, equally divided into 10 different concepts (classes). From each class, 40 images were used as queries and the remaining 40 images for retrieval. In the first set of experiments, the performance of SOD on the current database was assessed. The following 5 objects have been considered {African, bus, dinosaur, building, and flower}. A Receiver Operating Characteristic (ROC) curve plots the probability of correct detection of TPR (True Positive Rate) against the probability of false detection FPR (False Positive Rate). Figure 5.5 shows ROC curves for different objects. As can be seen, the performance demonstrates the efficiency of the proposed SOD technique. This also shows that SOD can be used to enhance CBIR.



(a)



(b)

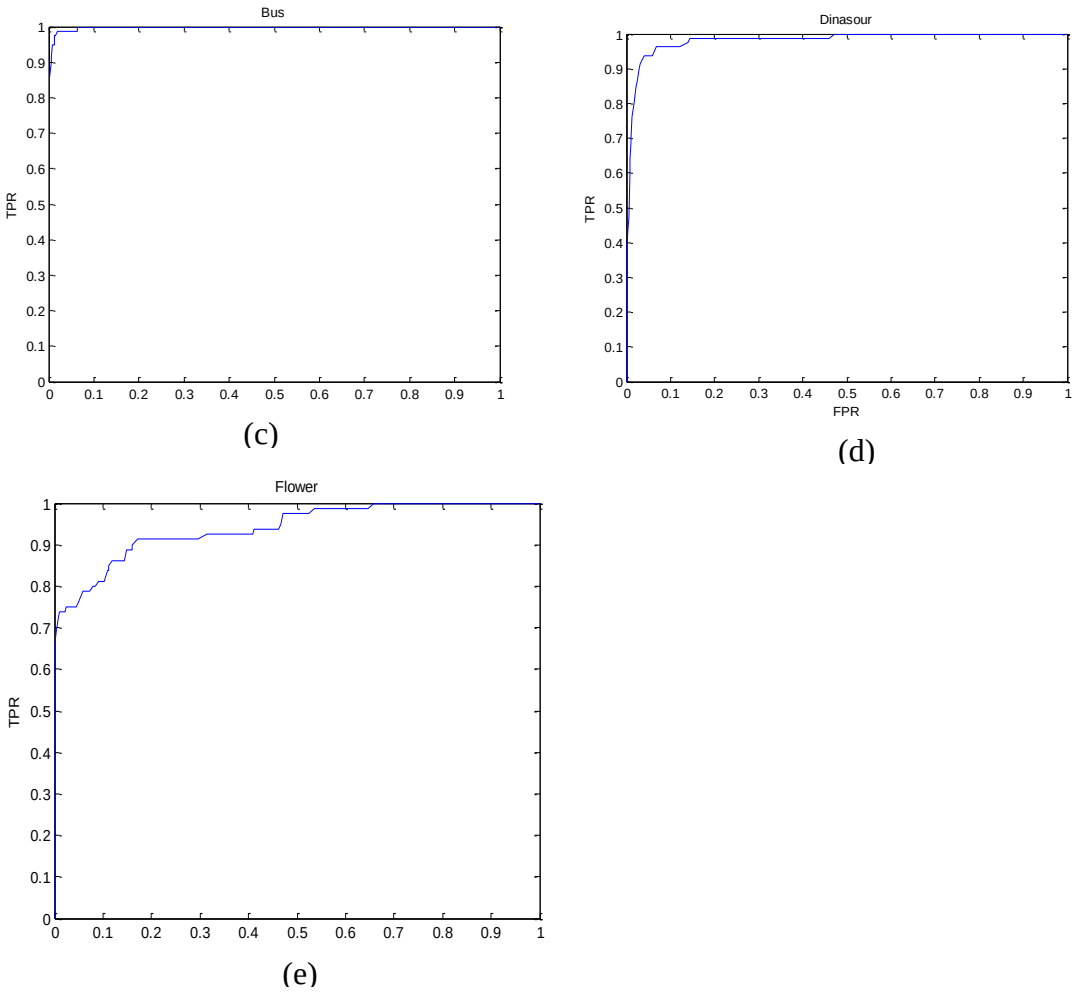


Figure 5-5 Semantic object detection results. (a) African (b) Building (c) Bus (d) Dinosaur (e) flower

Table 5.1 contains the retrieval results for various classes with different retrieval techniques. BDIP & BVLC refers to the CBIR technique proposed in [107] without SOD, while BDIP & BVLC refers to the CBIR technique proposed in Section 5.1.1 combined with SOD. As shown, the retrieval enhancements attributed to the use of SOD are also outperforms if the BDIP & BVLC only is taken as a reference point. Furthermore, the main contribution in the experiment results can be observed in the tables 5.1, 5.2, and 5.3, that the average retrieval rate of the proposed method is greater than or equal to the corresponding reference method.

Table 5.1 Average Retrieval rates with different techniques on WANG_DB

	BDIP&BVLC Ref. [8]				SOD_&_BDIP_BVLC (Proposed)			
CLASS	**Top 15**	**Top 20**	**Top 25**	**Top 30**	**Top 15**	**Top 20**	**Top 25**	**Top 30**
African	**55%**	**51.88%**	**48.60%**	**45.08%**	**55.33%**	**52.13%**	**49.40%**	**46.08%**
Beach	**43.17%**	**40.13%**	**37.5%**	**34.75%**	**43.17%**	**40.25%**	**37.6%**	**35%**
Building	**43.5%**	**39.88%**	**35.70%**	**33.17%**	**44%**	**40.38%**	**36.2%**	**34.08%**
Bus	**83.17%**	**78.75%**	**77.10%**	**74.92%**	**86.67%**	**82.25%**	**80.4%**	**78.92%**
Dinosaur	**99.5%**	**99.13%**	**98.8%**	**98.25%**	**99.5%**	**99.25%**	**99%**	**98.58%**
Elephant	**65.5%**	**61.38%**	**56.40%**	**53.5%**	**65.5%**	**61.38%**	**56.4%**	**53.5%**
Flower	**82.17%**	**79%**	**75.70%**	**73.75%**	**82.17%**	**79.13%**	**75.8%**	**74%**
Horse	**78.83%**	**73.75%**	**68.60%**	**64.5%**	**78.83%**	**73.75%**	**68.6%**	**64.58%**
Snow	**40.33%**	**36.75%**	**34.5%**	**32.75%**	**41%**	**37.88%**	**35.5%**	**33.83%**
Food	**48.67%**	**43.75%**	**40.9%**	**37.58%**	**49.17%**	**44.38%**	**41.4%**	**38.42%**
Average	**63.98%**	**60.44%**	**57.38%**	**54.83%**	**64.53%**	**61.08%**	**58.03%**	**55.7%**

In the second experiment, the Corel database [141] has been used to evaluate the performance of the proposed CBIR technique. 720 colour images have been used, equally divided into 9 different concepts (classes). From each class, 40 images were used as queries and the remaining 40 images for retrieval. In the experiments, the performance of SOD on the current database was assessed. The following 5 objects have been considered (Lion, Butterfly, Antique Car, horse, and Cambridge building). A Receiver Operating Characteristic (ROC) curve plots the probability of correct detection of TPR (True Positive Rate) against the probability of false detection FPR (False Positive Rate). Figure 5.7 shows ROC curves for different objects. As can be seen, the performance demonstrates the efficiency of the proposed SOD technique. This also shows that SOD can be used to enhance CBIR.

Table 5.2 and 5.3 depict the retrieval results for various classes with different retrieval techniques. BDIP & BVLC refers to the CBIR technique proposed in [107] without SOD while SOD and BDIP & BVLC refers to the CBIR technique proposed in section 5.1.1. As shown, the proposed scheme outperforms BDIP & BVLC techniques only. The retrieval enhancements attributed to the use of SOD are also noticeable compared with the different

top images of both BDIP&BVLC only and SOD_&_BDIP_BVLC if we take BDIP & BVLC as a reference point.

In addition, table 5.4 illustrated the retrieval results for various classes of Corel DB with different retrieval techniques. DCT only refers to the CBIR technique proposed in section 3.3.2.1, compared the DCT combined with Semantic object detection. As can be seen from the table, using semantic object detection combined with DCT enhancements the retrieval performance if we take DCT only as a reference point.

Table 5.2 . Average Retrieval rates with different techniques on 12 step block size Corel DB

	DIP&BVLC Ref. [8]				SOD_&_BDIP_BVLC (Proposed)			
CLASS	**Top 15**	**Top 20**	**Top 25**	**Top 30**	**Top 15**	**Top 20**	**Top 25**	**Top 30**
Lion	**54%**	**51.38%**	**50.30%**	**48.5%**	**54.67%**	**52.5%**	**51.20%**	**49.42%**
Sport car	**93.17%**	**91.5%**	**90.3%**	**88.92%**	**93.17%**	**91.5%**	**90.30%**	**89%**
Water	**52.83%**	**46.12%**	**41.90%**	**37.92%**	**53.17%**	**47.%**	**42.00%**	**38.08%**
Aircraft	**93.33%**	**92.75%**	**92%**	**90.75%**	**93.5%**	**92.88%**	**92%**	**90.92%**
Antique Car	**90%**	**89.12%**	**87.8%**	**85.92%**	**90.17%**	**89.25%**	**88%**	**86.17%**
Birds	**62.83%**	**60.88%**	**58.9%**	**57.5%**	**63.67%**	**61.63%**	**59.6%**	**58.08%**
Horse	**92.5%**	**89.5%**	**86.6%**	**84.17%**	**92.5%**	**89.75%**	**87%**	**84.75%**
Cambridge	**84.5%**	**82.25%**	**80.5%**	**78.25%**	**85.83%**	**84.25%**	**82.5%**	**80.5%**
Butterfly	**99.83%**	**99.88%**	**99.9%**	**99.92%**	**99.83%**	**99.88%**	**99.90%**	**99.92%**
Average	**80.33%**	**78.15%**	**76.47%**	**74.65%**	**80.72%**	**78.74%**	**76.94%**	**75.2%**

Table 5.3. Average Retrieval rates with different techniques on 8 step block size Corel DB

	BDIP&BVLC Ref. [8]				SOD_&_BDIP_BVLC (Proposed)			
CLASS	**Top 15**	**Top 20**	**Top 25**	**Top 30**	**Top 15**	**Top 20**	**Top 25**	**Top 30**
Lion	**54%**	**51.38%**	**50.30%**	**48.5%**	**54.67%**	**52.5%**	**51.20%**	**49.42%**
Sport car	**93.17%**	**91.5%**	**90.3%**	**88.92%**	**93.17%**	**91.5%**	**90.30%**	**89%**
Water	**52.83%**	**46.12%**	**41.90%**	**37.92%**	**53.17%**	**46.75%**	**42.00%**	**38.08%**
Aircraft	**93.33%**	**92.75%**	**92%**	**90.75%**	**93.5%**	**92.88%**	**92%**	**91%**
Antique Car	**90%**	**89.12%**	**87.8%**	**85.92%**	**90.17%**	**89.25%**	**88.10%**	**86.33%**
Birds	**62.83%**	**60.88%**	**58.9%**	**57.5%**	**63.17%**	**61.25%**	**59.40%**	**58%**
Horse	**92.5%**	**89.5%**	**86.6%**	**84.17%**	**92.67%**	**89.75%**	**86.80%**	**84.75%**
Cambridge	**84.5%**	**82.25%**	**80.5%**	**78.25%**	**85.5%**	**83.87%**	**82.3%**	**80.5%**
Butterfly	**99.83%**	**99.88%**	**99.9%**	**99.92%**	**99.83%**	**99.88%**	**99.90%**	**99.92%**
Average	**80.33%**	**78.15%**	**76.47%**	**74.65%**	**80.65%**	**78.63%**	**76.89%**	**75.22%**

Table 5.4 Average DCT and DCT with SOD Retrieval rates on 12 step block size Corel DB

	DCT Only				DCT & SOD (Proposed)			
CLASS	**Top 15**	**Top 20**	**Top 25**	**Top 30**	**Top 15**	**Top 20**	**Top 25**	**Top 30**
Lion	**46.83`%**	**43.63%**	**41.90%**	**38.42%**	**53.17%**	**49.63%**	**46.40%**	**44.17%**
Sport car	**46.17%**	**42.25%**	**40.10%**	**37.58%**	**53.17%**	**49.37%**	**46.30%**	**43.83%**
Water	**48%**	**43.63%**	**41.5%**	**39.42%**	**49.67%**	**45.87%**	**44.10%**	**41.75%**
Aircraft	**40.67%**	**37.37%**	**34.6%**	**33.25%**	**42.67%**	**40%**	**37.80%**	**37.17%**
Antique Car	**82%**	**80.63%**	**78.3%**	**75.92%**	**86.50%**	**85.75%**	**83.30%**	**81.25%**
Birds	**51.17%**	**48.5%**	**46.7%**	**43.42%**	**51.5%**	**49.13%**	**47.60%**	**44.08%**
Horse	**92.33%**	**88.87%**	**86.00%**	**82.92%**	**93.00%**	**90.12%**	**87.70%**	**85.17%**
Cambridge	**81.5%**	**80.5%**	**78.3%**	**75.92%**	**91.17%**	**89.50%**	**88.00%**	**85.83%**
Butterfly	**71.33%**	**66.38%**	**60.9%**	**57%**	**80.67%**	**76.12%**	**71.90%**	**70%**
Average	**62.22%**	**59.08%**	**56.48%**	**53.76%**	**66.83%**	**63.94%**	**61.46%**	**59.25%**

Another experiment applied on frontal face images to justify the face semantic image retrieval based on the semantic object detection, frontal face image was applied to the system detection. Figure 5.6 shows ROC curves for frontal face object among different objects. As can be seen, the performance indicates the probability of correct detection of TPR (True Positive Rate) against the probability of false detection FPR (False Positive Rate) and also demonstrates the efficiency of the proposed semantic object detection (SOD) technique. Table 5.5 depict the retrieval results for frontal face images among various classes with different retrieval techniques. BDIP & BVLC refers to the CBIR technique proposed in [107] without SOD while SOD and BDIP & BVLC refers to the CBIR technique proposed in section 5.1.1. As shown in the column of the frontal face images, the proposed scheme outperforms BDIP & BVLC techniques only. The retrieval enhancements attributed to the use of SOD are also noticeable compared with the different top images of both BDIP&BVLC only and SOD_&_BDIP_BVLC if we take BDIP & BVLC only as a reference point.

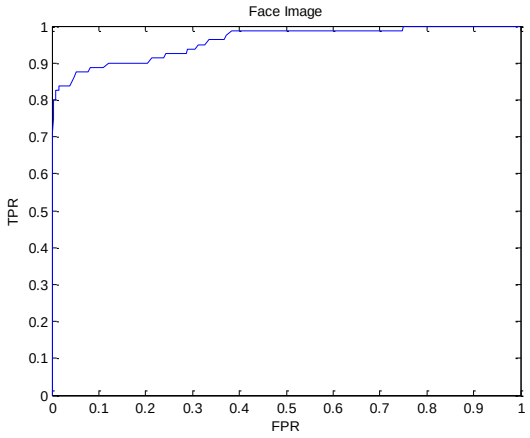


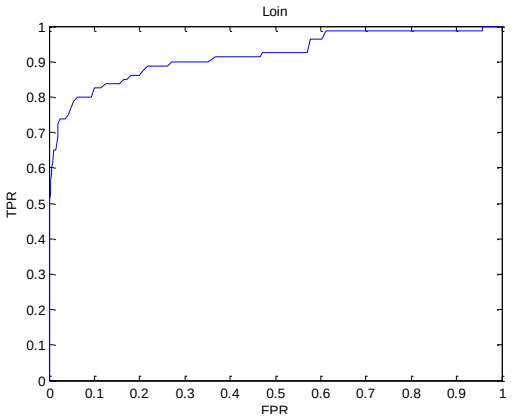
Figure 5-6 ROC curve of Frontal face image

Table 5.5 Retrieval restful of frontal face with different objects

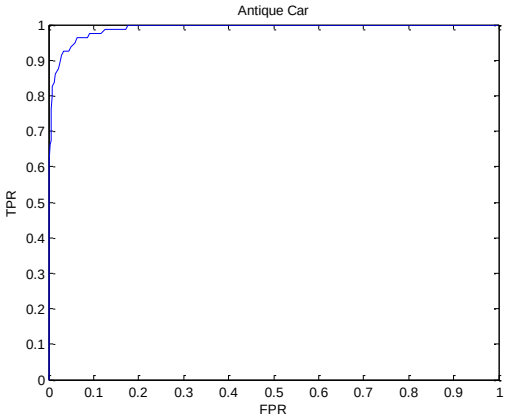
	BDIP&BVLC				SOD_&_BDIP_BVLC (Proposed)			
CLASS	**Top 15**	**Top 20**	**Top 25**	**Top 30**	**Top 15**	**Top 20**	**Top 25**	**Top 30**
Frontal Face	**70.50%**	**68.25%**	**66.80%**	**64.25%**	**70.67%**	**68.25%**	**66.80%**	**64.33%**
Beach	**41.50%**	**37.88%**	**35.20%**	**32.33%**	**41.50%**	**37.88%**	**35.40%**	**32.67%**
Building	**44.67%**	**40.75%**	**36.70%**	**34.00%**	**45.83%**	**41.63%**	**37.70%**	**35.33%**
Bus	**82.83%**	**79.75%**	**76.80%**	**74.50%**	**85.67%**	**82.25%**	**80.90%**	**78.67%**
Dinosaur	**99.5%**	**99.25%**	**98.8%**	**98.25%**	**99.5%**	**99.38%**	**98.90%**	**98.67%**
Elephant	**65.17%**	**61.13%**	**56.80%**	**53.17%**	**65.17%**	**61.13%**	**56.8%**	**53.17%**
Flower	**87.83%**	**85.38%**	**82.60%**	**80.50%**	**87.83%**	**85.50%**	**82.70%**	**80.67%**
Horse	**80.50%**	**75.38%**	**71.10%**	**66.83%**	**80.50%**	**75.38%**	**71.10%**	**66.92%**
Snow	**40.17%**	**35.50%**	**33.70%**	**31.75%**	**40.50%**	**36.25%**	**34.00%**	**32.33%**
Food	**49.00%**	**44.13%**	**41.40%**	**38.00%**	**49.17%**	**44.87%**	**42.00%**	**38.75%**
Average	**66.17%**	**62.74%**	**59.99%**	**57.35%**	**66.64%**	**63.25%**	**60.63%**	**58.15%**

5.4 **Conclusion**

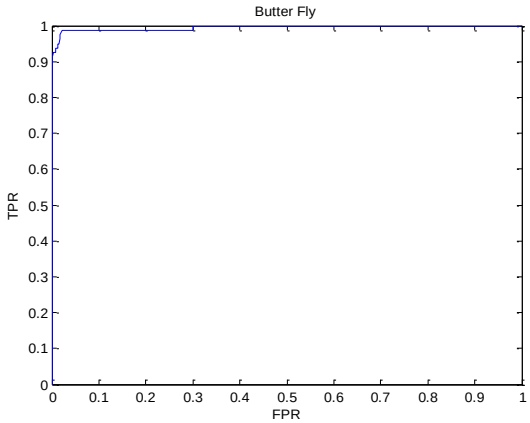
In the proposed work described earlier is an efficient content based image retrieval technique based on semantic object detection. The idea relies on the fact that existing CBIR schemes could be improved by using further high level knowledge to filter the database for retrieval of similar images. It has been shown through experimental results that the proposed scheme (semantic object detection) outperforms both BDIP&BVLC and DCT retrieval technique and another related technique from the literature.



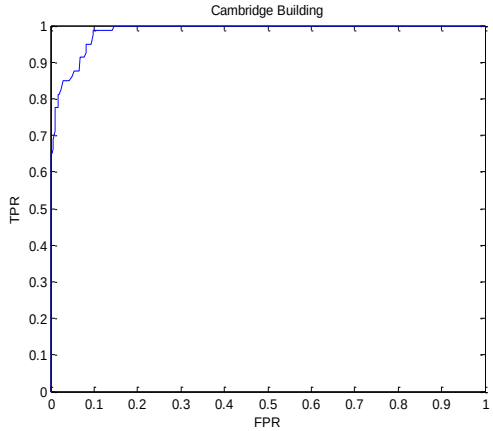
(a)



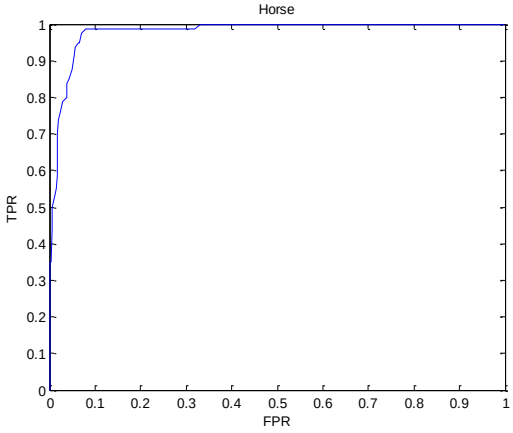
(b)



(c)



(d)



(e)

Figure 5-7 Semantic object detection results. (a) Loin (b) Antique Car (c) Butterfly (d) Cambridge Building (e) Horse

Chapter SIX

6 Conclusion and future work

The work presented in this thesis is an initial contribution towards a future CBIR system in order to narrow the gap between low-level features and semantic interpretation. More than twenty years since CBIR became a major activity in the image processing. CBIR Research is carried out in the DCT domain. It is still difficult to evaluate how successful content-based image retrieval systems are, in terms of effectiveness and efficiency. Combining visual features has become a more recent focus of recent researches and developing system that narrow the gap between high level and low level features in order to improve the accuracy of the image retrieval.

CBIR in the DCT domain has become more necessary as most images are compressed at the JPEG source. There have been much works published on how to increase the effectiveness and efficiency of content based retrieval in the DCT domain over the past few years, but the results are still far from the expectations.

The major work done in this thesis is presented in chapters 3, 4 and 5.

Chapter 3 discusses the extraction of DCT coefficients to construct an efficient DCT histogram quantization method for image retrieval and image classification. The extraction method directly from DCT transform domain has the capability of producing high average precision in image retrieval and image classification as shown by the proposed method. It also can be stated that the DC and some of the AC's coefficients extraction method only needs a portion of the time needed for the indexing process compared to using all the DCT coefficients.

In addition, different feature extraction was investigated for content based image retrieval such as image sub-blocks, non-overlapping, and DCT wavelet sub-bands.

Chapter 4 presents low level extraction based on face detection using skin face colour, where colour feature is widely used as an effective feature in the face detection algorithms. The threshold based, Gaussian mixture model based descriptors have been used in the DCT domain to improve detection and classification of face images using neural networks.

Chapter 4 also discusses the extraction a feature directly from DCT coefficients to construct a simple, efficient and effective method for face detection. The low frequency coefficients extraction method has the capability of producing high detection rate in frontal face image detection and decreases the amount of storage required. It also can be stated that the first low frequency coefficients extraction method only needs a portion of the time needed for the indexing process compared with using all the DCT coefficients.

In Chapter 5, combinations of hybrid methods have been proposed to improve the effectiveness of content based image retrieval. In order to bridge the gap between the low level and the high level feature (semantic), a new content-based image retrieval scheme based on Semantic Object Detection (SOD) was proposed. The feature extraction process uses quantized discrete cosine transform (DCT) blocks and the retrieval of the query image is performed using a histogram-based similarity measure. SOD aims to reduce the size of the database from which the retrieval of similar images is conducted. The use of SOD has been shown to improve the retrieval performance and narrow the semantic gab between the low level and the high level feature.

6.1 Thesis Contributions

The main objective of this thesis is to classify and retrieve query images from image databases. The existing techniques were studied in terms of pre-processing, segmentation, feature extraction, detection, classification and retrieval. In this thesis, the contributions can be summarized as follows:

(1) DCT features descriptors are presented to improve the effectiveness of face detection and the content based image retrieval in the DCT domain. Skin colour thresholds and likelihood based Gaussian mixture models have been used to asses classification of face and non-face using features extracted directly from DCT coefficients. The result of face and non-face recognition achieved significant result as described in section 4.1.4 in chapter 4. Meanwhile, content based face image retrieval was presented and reasonable results were achieved.

(2) A selection the low frequency coefficients instead of the whole DCT coefficients simplify the cost of the calculation and reduce the space of the storage.

(3) Normalizing the selected DCT coefficient by calculating the min and max values of the coefficient asses to achieve a good representation of the image features and resolve the sensitivity to the image operation (brightness, contrast, histogram equalization)

(4) Labelling the output of Neural Networks using the average of the input features and binary numbers which produced good results in the face classification and in the in the image retrieval.

(5) In order to simplify calculations and improve precision, the DC and some of AC's extraction approach was used in image indexing and retrieval. Instead of working

on DCT images, this extraction method worked on a portion of image blocks only to represent DCT images.

(6) DCT histogram Quantization feature extraction from a few DCT coefficients as a feature vectors to represent the content of the input image. In order to count the number of coefficients having the same DCT coefficient over all image blocks. Significant image classification results have been achieved compared with the literature as shown in chapter 3.

(7) In order to narrow the semantic gab between the low level and high level features semantic object detection (SOD) and neural networks was introduced in chapter 3 and 5. SOD exploits the idea that images could be described by high level knowledge-based concepts which in turn can be represented by specific objects. Semantic object detection (SOD) based image retrieval was presented in Chapter 5,

(8) Using SOD by applying object detection classification reduce the database compared with the test query image by excluding non object database from the original database image.

(9) Applicable to any CBIR algorithms, the proposed technique SOD is combined to works with DCT only and another re-implemented adopted technique BDIP& BVLC and good outperformance results achieved as described in chapter 5.

6.3. Future Work

This thesis in itself does not attempt to solve all the problems faced by the CBIR community, there still other areas to be explored as possible future work as indicated below.

1. Besides using JPEG images which mainly rely on the discrete cosine transform, further investigation should consider images compressed in the JPEG2000 format where The wavelet transform has been adopted by MPEG-4 for still image coding (mpeg4). Also, JPEG-2000 is considering using the wavelet transform as its core technique for the next generation of the still image coding standard (jpeg2000). This is because the wavelet transform can provide not only excellent coding efficiency, but also good spatial and quality scalable functionality. JPEG-2000 is a new type of image compression system under development by Joint Photographic Experts Group for still image coding., since wavelet transformation are good at describing image texture.
2. Implement features suitable for other classifiers such as support vector machines (SVMs). These are mainly used for binary classifications and are capable of generating fast classifier functions following a training period. There are several techniques for adapting SVMs to multi-class classification problems with three or more classes. SVM light is a multi-class classifier which is available at http://svmlight.joachims.org/svm_multiclass.html.
3. Ontology can be used not only for annotation and precise information retrieval [142], but also for helping the user in formulating the information needed and the corresponding query. In this case, the ontology has an enriched knowledge base of image metadata which can be applied to construct more meaningful retrieval rather than just by description or annotation about images. By using this enriched knowledge, image ontology can provide semantic browsing facilities. The most

difficult problem that might be considered for the future work in building image ontology is how to automate annotation with high accuracy.

4. Another subject to be explored is relevance feedback (RF). This is a supervised learning technique which is used to improve the effectiveness of information retrieval systems. It was introduced to CBIR during the mid 1990s with the intention to incorporate the user in the retrieval loop in order to reduce the semantic gap between what the query represents (low level features) and what the user think.
5. Automated annotation is widely recognized as an extremely difficult issue.Word picture or text image approach is significant direction of treating the problem of image annotation by integrate textual with visual data image.

7 References

1. Faloutsos, C., et al., Efficient and effective querying by image content. Journal of intelligent information systems, 1994. **3**(3): p. 231-262.

2. Mukherjea, S., K. Hirata, and Y. Hara, AMORE: A World Wide Web image retrieval engine. World Wide Web, 1999. **2**(3): p. 115-132.

3. Pentland, A., R.W. Picard, and S. Sclaroff, Photobook: Content-based manipulation of image databases. International Journal of Computer Vision, 1996. **18**(3): p. 233-254.

4. Smith, J.R. and S.F. Chang. VisualSEEk: a fully automated content-based image query system. 1997: ACM.

5. Wang, J.Z., J. Li, and G. Wiederhold, SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on pattern analysis and machine intelligence, 2001: p. 947-963.

6. Del Bimbo, A., Visual information retrieval. 1999: Morgan Kaufmann.

7. Shneier, M. and M. Abdel-Mottaleb, Exploiting the JPEG compression scheme for image retrieval. IEEE Transactions on pattern analysis and machine intelligence, 1996. **18**(8): p. 849-853.

8. Le Gall, D., MPEG: A video compression standard for multimedia applications. Communications of the ACM, 1991. **34**(4): p. 46-58.

9. Feng, G. and J. Jiang, JPEG compressed image retrieval via statistical features. Pattern Recognition, 2003. **36**(4): p. 977-985.

10. Datta, R., J. Li, and J.Z. Wang. Content-based image retrieval: approaches and trends of the new age. 2005: ACM New York, NY, USA.

11. Mandal, M., F. Idris, and S. Panchanathan, A critical evaluation of image and video indexing techniques in the compressed domain. Image and Vision Computing, 1999. **17**(7): p. 513-529.

12. Shen, B. and I. Sethi. Direct feature extraction from compressed images. proceeding in SPIE v.2670 1996:pp.404-414.

13. Reeves, R., K. Kubik, and W. Osberger. Texture characterization of compressed aerial images using DCT coefficients. Proceeding in SPIE Vol. 3022, p. 398-407 1997.

14. C.-W. Ngo, T.-C. Pong, and R.T. Chin, Exploiting image indexing techniques in DCT domain. Pattern Recognition, 2001. **34**(9): p. 1841–1845.

15. Sethi, I., I. Coman, and D. Stan, Mining association rules between low-level image features and high-level concepts. Proceedings of the SPIE Data Mining and Knowledge Discovery, 2001. **3**: p. 279–290.

16. Mojsilovic, A. and B. Rogowitz. Capturing image semantics with low-level descriptors. international conference on image processing vol.1.2001.pp 18-21.

17. Zhou, X. and T. Huang. CBIR: from low-level features to high-level semantics. 2000: Image and video communications and processing.2000.pp.25-28.

18. Hjelmås, E. and B. Low, Face detection: A survey. Computer Vision and Image Understanding, 2001. **83**(3): p. 236-274.

19. Sakai, T., M. Nagao, and T. Kanade. Computer analysis and classification of photographs of human faces. 1972.

20. Luo, H. and A. Eleftheriadis. On face detection in the compressed domain. 2000: ACM New York, NY, USA.

21. Wang, J.,Mohan K, Abdulredha.H,. Face detection using DCT coefficients in MPEG video: In proceedings of International Workshop on Advanced Image Technology.2002.

22. Heisele, B., T. Poggio, and M. Pontil, Face detection in still gray images. 2000.

23. Wang, H., H. Stone, and S. Chang, FaceTrack: Tracking and summarizing faces from compressed video. SPIE Multimedia Storage and Archiving Systems IV, Boston, 1999.

24. Sobottka, K. and I. Pitas. Extraction of facial regions and features using color and shape information. 1996: Citeseer.

25. Terrillon, J., Fukamachi .H,Akamatsu.S,. Shirazi M. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. 2000: IEEE Computer Society Washington, DC, USA.

26. Govindaraju, V., Locating human faces in photographs. International Journal of Computer Vision, 1996. **19**(2): p. 129-146.

27. Yang, M., D. Kriegman, and N. Ahuja, Detecting faces in images: A survey. IEEE Transactions on pattern analysis and machine intelligence, 2002: p. 34-58.

28. Yang, M. and N. Ahuja, Detecting human faces in color images. Urbana. **51**: p. 61801.

29. Oliver, N., A. Pentland, and F. Bérard, LAFTER: a real-time face and lips tracker with facial expression recognition. Pattern Recognition, 2000. **33**(8): p. 1369-1382.

30. Yang, J. and A. Waibel. A real-time face tracker. 1996: Citeseer.

31. Sakai, T., M. Nagao, and S. Fujibayashi, Line extraction and pattern detection in a photograph. Pattern Recognition, 1969. **1**(3): p. 233-236.

32. Lanitis, A., C. Taylor, and T. Cootes, Automatic face identification system using flexible appearance models. Image and Vision Computing, 1995. **13**(5): p. 393-401.

33. Yang, G. and T. Huang, Human face detection in a complex background. Pattern Recognition, 1994. **27**(1): p. 53-63.

34. Rowley, H., S. Baluja, and T. Kanade, Neural network-based face detection. IEEE Transactions on pattern analysis and machine intelligence, 1998. **20**(1): p. 23-38.

35. Turk, M. and A. Pentland, Eigenfaces for recognition. Journal of cognitive neuroscience, 1991. **3**(1): p. 71-86.

36. Sung, K. and T. Poggio, Example-based learning for view-based human face detection. IEEE Transactions on pattern analysis and machine intelligence, 1998. **20**(1): p. 39-51.

37. Yuille, A., P. Hallinan, and D. Cohen, Feature extraction from faces using deformable templates. International Journal of Computer Vision, 1992. **8**(2): p. 99-111.

38. Papageorgiou, C., M. Oren, and T. Poggio, A general framework for object detection. Sixth international conference on computer vision .1998.pp 555-562

39. Podilchuk, C. and X. Zhang, Face recognition using DCT-based feature vectors. proceeding of Acoustics, speech,and signal processing 1998, pp.2144-2147.

40. Wang, H. and S. Chang, A highly efficient system for automatic face region detection in MPEG video. IEEE Transactions on Circuits and Systems for Video Technology, 1997. **7**(4): p. 615-628.

41. Garcia, C. and G. Tziritas, Face detection using quantized skin color regions merging andwavelet packet analysis. IEEE Transactions on Multimedia, 1999. **1**(3): p. 264-277.

42. Rowley;, H., S. Baluja;, and T. Kanade, Rotation Invariant Neural Network-Based Face Detection. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 1998.pp.38-44

43. Mehrotra, S., et al. Supporting content-based queries over images in MARS. 1997.

44. Gevers, T. and A. Smeulders, Pictoseek: Combining color and shape invariant features for image retrieval. IEEE Transactions on Image Processing, 2000. **9**(1): p. 102-119.

45. Agui, T., et al., Extraction of FaceRecognition from monochromatic photographs using neural networks. 1992. p. 1881–1885.

46. Fahlmann, S.E., Lebiere, C., Advances in Neural Information Processing System 2(NIPS-2), in Touretzky, D.S. 1989: Morgan Kaufmann, Denver. p. 524.

47. Miao, J., et al., A hierarchical multiscale and multiangle system for human face detection in a complex background using gravity-center template. Pattern Recognition, 1999. **32**(7): p. 1237-1248.

48. Shet, R.N., Lai, K.H., Edirisingh, E., Chung, P.W.H. Pattren Recognition and Image Analysis , Lecture Notes in Computer Science. in Marques,J.S.,de la Pe'rez,B.N.,Pina,P. 2005. Berlin: Springer.

49. DTREG, http://www.dtreg.com/cascade.htm.

50. Farah, N., L. Souici, and M. Sellami, Classifiers combination and syntax analysis for Arabic literal amount recognition. Engineering Applications of Artificial Intelligence, 2006. **19**(1): p. 29-39.

51. Fukunaga, K., Introduction to Statistical Pattern Recognition, in Academic. 1990: New York, USA. p. 220.

52. Fawcett, T., An introduction to ROC analysis. Pattern Recognition Letters, ElsevieR, 2006. **27**: p. 861-874.

53. Chernesky, M., Jang, D., Krepel, J., Sellors, J. & Mahony, J., Impact of reference standard sensitivity on accuracy of rapid antigen detection assays and a leukocyte esterase dipstick for diagnosis of Chlamydia trachomatis infection in first-void urine specimens from men. Journal of Clinical Microbiology 1999. **37**: p. 2777 - 2780.

54. Wallace, G., The JPEG still picture compression standard. 1991.ACM .New york Ny.USA

55. Ahmed, N., T. Natarajan, and K. Rao, Discrete cosine transfom. IEEE Transactions on Computers, 1974. **100**(23): p. 90-93.

56. Pratt, W., Digital image processing: PIKS inside. 2001: John Wiley & Sons, Inc. New York, NY, USA.

57. Lee, S., H. Bae, and S. Jung, Efficient content-based image retrieval methods using color and texture. ETRI journal, 1998. **20**(3): p. 272-283.

58. Chen, W. and W. Pratt, Scene adaptive coder. Communications, IEEE Transactions on [legacy, pre-1988], 1984. **32**(3): p. 225-232.

59. Pittner, S. and S.V. Kamarthi, Feature extraction from wavelet coefficients for pattern recognition tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999. **21**(1): p. 83-88.

60. Ferreira, C.B.R. and D.L. Borges. Automated mammogram classification using a multiresolution pattern recognition approach. in Computer Graphics and Image Processing, 2001 Proceedings of XIV Brazilian Symposium on. 2001.

61. Faye, I., B.B. Samir, and M.M. Eltoukhy. Digital Mammograms Classification Using a Wavelet Based Feature Extraction Method. in Second International Conference on Computer and Electrical Engineering, 2009. ICCEE '09. . 2009.

62. Ritendra, D., et al., Image retrieval: ideas, influences, and trends of the new age. ACM Computing Surveys, 2005. **40**(2): p. 1-60.

63. Gupta, A. and R. Jain, Visual information retrieval. Communications of the ACM, 1997. **40**(5): p. 70-79.

64. Carson, C., et al., Blobworld: A system for region-based image indexing and retrieval. Lecture Notes in Computer Science, 1999: p. 509-516.

65. Zhang, Y., Toward High-Level Visual Information Retrieval. Semantic-based visual information retrieval, 2006: pp. 1.22

66. Squire, D., et al., Content-based query of image databases: inspirations from text retrieval. Pattern Recognition Letters, 2000. **21**(13-14): p. 1193-1198.

67. Inoue, M. and N. Ueda. Retrieving lightly annotated images using image similarities. Proceeding of the ACM symposium on Applied computing .2005.pp 1031-1037

68. Inoue, M. On the need for annotation-based image retrieval. 2004.

69. Yang, Z. and C. Kuo, Survey on image content analysis, indexing, and retrieval techniques and status report of MPEG-7. Tamkang Journal of Science and Engineering, 1999. **2**(3): p. 101-118.

70. Mandal, M., F. Idris, and S. Panchanathan. Image and video indexing in the compressed domain. 1997.

71. Feig, E. and S. Winograd, Fast algorithms for the discrete cosine transform. IEEE Transactions on Signal Processing, 1992. **40**(9): p. 2174-2193.

72. Rui, Y., T.S. Huang, and S.F. Chang, Image retrieval: Current techniques, promising directions, and open issues. Journal of visual communication and image representation, 1999. **10**(1): p. 39-62.

73. Swain, M. and D. Ballard, Color indexing. International Journal of Computer Vision, 1991. **7**(1): p. 11-32.

74. Smith, J. and S. Chang, Automated binary texture feature sets for image retrieval. In Proceedings of the IEEE International Conference Acoustics, Speech, and Signal Processing, 1996.pp 2239-2242

75. Haralick, R., I. Dinstein, and K. Shanmugam, Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, 1973. 3(6): p. 610-621.

76. Tamura, H., S. Mori, and T. Yamawaki, Textural features corresponding to visual perception. IEEE Transactions on Systems, Man and Cybernetics, 1978. **8**(6): p. 460-473.

77. Cohen, F., Texture classification by wavelet packet signatures. IEEE Transactions on pattern analysis and machine intelligence, 1993. **15**(11).

78. Smith, J. and S. Chang. Transform features for texture classification and discrimination in large image databases. 1994.

79. Mehtre, B., M. Kankanhalli, and W. Lee, Shape measures for content based image retrieval: a comparison. Information processing and Management, 1997. **33**(3): p. 319-337.

80. Jose, J., J. Furner, and D. Harper. Spatial querying for image retrieval: a user-oriented evaluation. 1998: ACM New York, NY, USA.

81. Iqbal, Q. and J. Aggarwal, Image retrieval via isotropic and anisotropic mappings. Pattern Recognition, 2002. **35**(12): p. 2673-2686.

82. Provost, F., T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. Proceedings of the Fifteenth International Conference on Machine Learning **1**998: pp.445-453.

83. Mandal, M., T. Aboulnasr, and S. Panchanathan, Image indexing using moments and wavelets. IEEE Transactions on Consumer Electronics, 1996. **42**(3): p. 557-565.

84. Bhaskaran, V. and K. Konstantinides, Image and video compression standards: algorithms and architectures. 1997: Kluwer Academic Pub.

85. Feng, G. and J. Jiang, Image extraction in DCT domain. IEE Proceedings-Vision, Image, and Signal Processing, 2003. **150**: p. 20-27

86. Mojsilovic, A. and B. Rogowitz. Capturing image semantics with low-level descriptors: Proceeding of international conference of image processing. 2001. pp 18-21.

87. Liu, Y., D.Zhang,G.Lu,W.Ma., A survey of content-based image retrieval with high-level semantics. Pattern Recognition, 2007. **40**(1): p. 262-282.

88. Long, F., H. Zhang, and D. Feng, Fundamentals of content-based image retrieval. Multimedia Information Retrieval and Management-Technological Fundamentals and Applications, D. Feng, WC Siu, and HJ Zhang (Eds.), Springer, 2003.

89. Eakins, J. Automatic image content retrieval-are we getting anywhere? 1996: Citeseer.

90. Mezaris, V., I. Kompatsiaris, and M.G. Strintzis, An ontology approach to object-based image retrieval, in Proceedings of the ICIP. 2003. p. 511-514.

91. H.Feng and T.S.Chua, "A boosting approach to annotating large image collection" Workshop on Multimedia Information Retrieval in ACM Multimedia, DEC 2004: p. 931-938.

92. Jin, W., R. Shi, and T.S. Chua, A semi-naïve Bayesian method incorporating clustering with pair-wise constraints for auto image annotation. 2004, ACM. p. 339.

93. Chang, S., W. Chen, and H. Sundaram. Semantic visual templates: linking visual features to semantics. 1998.

94. Rui, Y., T. Huang, and S. Chang, Image Retrieval: Current Techniques, Promising Directions, and Open Issues* 1. Journal of visual communication and image representation, 1999. **10**(1): p. 39-62.

95. Deb, S. and Y. Zhang. An overview of content-based image retrieval techniques. 2004.

96. Jiang, J., A. Armstrong, and G. Feng, Direct content access and extraction from JPEG compressed images. Pattern Recognition, 2002. **35**(11): p. 2511-2519.

97. Vadivel, A., A. Majumdar, and S. Sural, Characteristics of weighted feature vector in content-based image retrieval applications. Intelligent Sensing and Information Processing, 2004. **1**(18): p. 127-132.

98. Carson, C., et al. Blobworld: A system for region-based image indexing and retrieval: Springer.

99. Z.Z.Wang and J.H.Yong, Texture analysis and classification with linear regression model based on wavelet transform. IEEE Transaction . On Image Processing,, Aug.2008. **17**: p. 1421-1430.

100. http://wang.ist.psu.edu/docs/related.shtml:.

101. Rao, A., R. Srihari, and Z. Zhang. Spatial color histograms for content-based image retrieval. 1999: Citeseer.

102. Aamer Mohamed , et al., An Efficient Feature Extraction Technique based on the histogram of Block DCT. Proceeding of Signal Processing, Pattern Recognition and Applications Innsbruck, Austria 2010.

103. Smeulders, A., et al., Content-based image retrieval at the end of the early years. IEEE Transactions on pattern analysis and machine intelligence, 2000. **22**(12): p. 1349-1380.

104. Swain, M. and D. Ballard. Indexing via color histograms. 1990.

105. Mehtre, B. and M. Kankanhalli, Shape measures for content based image retrieval: a comparison. Information Processing & Management, 1997. **33**(3): p. 319-337.

106. Manjunath, B. and W. Ma, Texture features for browsing and retrieval of image data. IEEE Transactions on pattern analysis and machine intelligence, 1996. **18**(8): p. 837-842.

107. Chun, Y., S. Seo, and N. Kim, Image retrieval using BDIP and BVLC moments. IEEE Transactions on Circuits and Systems for Video Technology, 2003. **13**(9): p. 951-957.

108. Chun, Y., N. Kim, and I. Jang, Content-based image retrieval using multiresolution color and texture features. IEEE Transactions on Multimedia, 2008. **10**(6): p. 1073-1084.

109. Mezaris, V., I. Kompatsiaris, and M. Strintzis. An ontology approach to object-based image retrieval. 2003.

110. Vailaya, A., A. Jain, and H.J. Zhang, On image classification: City images vs. landscapes. 1998, Pattern Recognition. p. 1921-1935.

111. Schneiderman, H. and T. Kanade, Object detection using the statistics of parts, in International Journal of Computer Vision. 2004. p. 151-177.

112. Chua, T.S., Y. Zhao, and M.S. Kankanhalli, Detection of human faces in a compressed domain for video stratification, in The Visual Computer. 2002. p. 121-133.

113. Lew, M.S., Next-generation web searches for visual content. 2000, IEEE Computer Society. p. 46-53.

114. Fan, J., Y. Gao, and H. Luo, Multi-level annotation of natural scenes using dominant image components and semantic concepts, in Proceedings of the ACM International Conference on Multimedia. 2004. p. 540-547.

115. Li, J. and J.Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, in IEEE Transactions on Pattern Analysis and Machine Intelligence. 2003. p. 1075-1088.

116. Rautiainen, M., et al., Detecting semantic concepts from video using temporal gradients and audio classification, in Proceedings of the 3rd International Conference on Image and Video Retrieval. p. 397-402.

117. Amir, A., et al., A multi-modal system for the retrieval of semantic video events, in Computer Vision and Image Understanding. 2004. p. 216-236.

118. Li, J., J. Wang, and G. Wiederhold. IRM: integrated region matching for image retrieval. in Proceedings of the eighth 2000: ACM.

119. Rubner, Y., L. Guibas, and C. Tomasi. The earth mover's distance, multi-dimensional scaling, and color-based image retrieval. in Proceedings of the ARPA Image. 1997:

.

120. Banerjee, M., M. Kundu, and P. Das. Image Retrieval with Visually Prominent Features using Fuzzy set theoretic Evaluation. IET International Conference on visual information engineering.2006.pp.298-303

121. Hiremath, P. and J. Pujari, Content Based Image Retrieval using Color Boosted Salient Points and Shape features of an image. International Journal of Image Processing, 2008. **2**(1): p. 10-17.

122. Sheng, Y., A.H. Sadka, and A.M. Kondoz, Automatic 3D face synthesis using single 2D video frame, in Electronics Letters. 2004. p. 1173-1175.

123. Saber, E. and A. Tekalp, Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. Pattern Recognition Letters, 1998. **19**(8): p. 669-680.

124. Wang, J. and T. Tan, A new face detection method based on shape information. Pattern Recognition Letters, 2000. **21**(6-7): p. 463-471.

125. Moghaddam, B. and A. Pentland, Probabilistic visual learning for object representation. IEEE Transactions on pattern analysis and machine intelligence, 1997. **19**(7): p. 696-710.

126. Osuna, E., R. Freund, and F. Girosit. Training support vector machines: an application to face detection. 1997.

127. Chua, T., Y. Zhao, and M. Kankanhalli, Detection of human faces in a compressed domain for video stratification. The Visual Computer, 2002. **18**(2): p. 121-133.

128. Graf, H., et al. Locating faces and facial parts. 1995.

129. Graf, H., et al. Multi-modal system for locating heads and faces. 1996.

130. Lin, H., et al. Face detection based on skin color segmentation and neural network. 2005.

131. Jiang, J., Y. Weng, and P. Li, Dominant colour extraction in DCT domain. Image and Vision Computing, 2006. **24**(12): p. 1269-1277.

132. Ma, L., et al. A new facial expression recognition technique using 2D DCT and k-means algorithm. 2004.

133. Wong, K., K. Lam, and W. Siu. An efficient color compensation scheme for skin color segmentation. 2003.

134. Cho, K., J. Jang, and K. Hong, Adaptive skin-color filter. Pattern Recognition, 2001. 34 (5): p. 1067-1073.

135. Mohamed, A., Weng.Y, Ipson.S. Jiang.J.Face detection based on skin color in image by neural networks. International and advanced systems 2007.pp.779-78

136. Mostafa, L. and S. Abdelazeem. Face Detection Based on Skin Color Using Neural Networks. International journal on graphics, vision and image processing (GVIP) 2007.

137. Haralick, R., K. Shanmugam, and I. Dinstein, Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics, 1973. **3**(6): p. 610-621.

138. Manjunath, B., et al., Color and texture descriptors. IEEE Transactions on Circuits and Systems for Video Technology, 2001. **11**(6): p. 703-715.

139. Smith, J. and S. Chang. Transform features for texture classification and discrimination inlarge image databases. 1994.

140. Manjunath, B., P. Salembier, and T. Sikora, Introduction to MPEG-7: multimedia content description interface. 2002: John Wiley & Sons Inc.

141. http://vision.stanford.edu/resources_links.html.

142. Schreiber, G., et al., A mini-experiment in semantic annotation. The Semantic Web—ISWC 2002, 2002: p. 404-408.